



Volume 105
Number 4

November 2013

Published quarterly
by the
American Psychological
Association

ISSN 0022-0663
ISBN-10: 1-4338-1713-6
ISBN-13: 978-14338-1713-7

Journal of Educational Psychology

Special Issue: Advanced Learning Technologies

Guest Editors: Vincent Alevén, Carole R. Beal, and Arthur C. Graesser

Arthur C. Graesser, *Editor*

Jill Fitzgerald, *Associate Editor*

David Francis, *Associate Editor*

Susan Goldman, *Associate Editor*

Young-Suk Kim, *Associate Editor*

Robert Klassen, *Associate Editor*

David N. Rapp, *Associate Editor*

Susan Sonnenschein, *Associate Editor*

Birgit Spinath, *Associate Editor*

Roman Taraban, *Associate Editor*

Jennifer Wiley, *Associate Editor*

Christopher A. Wolters, *Associate Editor*

www.apa.org/pubs/journals/edu

Marygrove College Library
8425 W. McNichols Rd.
Detroit, MI 48221-2599

Editor

Arthur C. Graesser, *University of Memphis*

Associate Editors

Jill Fitzgerald, *University of North Carolina at Chapel Hill, Emeritus*
David Francis, *University of Houston*
Susan Goldman, *University of Illinois, Chicago*
Young-Suk Kim, *Florida State University*
Robert Klassen, *The University of York, United Kingdom*
David N. Rapp, *Northwestern University*
Susan Sonnenschein, *University of Maryland*
Birgit Spinath, *University of Heidelberg, Heidelberg, Germany*
Roman Taraban, *Texas Tech University*
Jennifer Wiley, *University of Illinois at Chicago*
Christopher A. Wolters, *The Ohio State University*

Chief Editorial Assistant

Jean Edgar, *University of Memphis*

Advisory Editors

Mary D. Ainley, *University of Melbourne, Australia*
Vincent Alevén, *Carnegie Mellon University*
Patricia Alexander, *University of Maryland, College Park*
Richard L. Allington, *University of Tennessee*
Ellen R. Altermatt, *Hanover College*
Ivan Ash, *Old Dominion University*
Carole Beal, *University of Arizona*
Hefer Bembenutty, *Queens College, CUNY*
David A. Bergin, *University of Missouri, Columbia*
Daniel Bolt, *University of Wisconsin, Madison*
Mimi Bong, *Ewha Womans University, Seoul, Korea*
Julie L. Booth, *Temple University*
Lee Brannum-Martin, *Georgia State University*
M. Anne Britt, *Northern Illinois University*
Scott Brown, *University of Connecticut*
Eric S. Buhs, *University of Nebraska, Lincoln*
Adriana G. Bus, *Leiden University, The Netherlands*
Kirsten R. Butcher, *University of Utah*
Robert Calfee, *University of California, Riverside*
Martha Carr, *University of Georgia*
Kwansu Cho, *University of Missouri, Columbia*
Timothy Cleary, *University of Wisconsin, Milwaukee*
Anne E. Cook, *University of Utah*
Kai Cortina, *University of Michigan*
Jennifer Cromley, *Temple University*
H. Michael Crowson, *University of Oklahoma*
Anne E. Cunningham, *University of California, Berkeley*
Teresa K. DeBacker, *The University of Oklahoma*
Sidney D'Mello, *University of Notre Dame*
John Dunlosky, *Kent State University*
Amanda M. Durik, *Northern Illinois University*
Gary Feng, *Educational Testing Service*
J. D. Fletcher, *Institute for Defense Analyses*
Lynn S. Fuchs, *Vanderbilt University*
Linda Gambrell, *Clemson University*
James P. Gee, *Arizona State University*
Arthur M. Glenberg, *Arizona State University*
Adele E. Gottfried, *California State University*
Steve Graham, *Arizona State University*
Barbara A. Greene, *University of Oklahoma*
John Guthrie, *University of Maryland*
Douglas Hacker, *University of Utah*
Vernon C. Hall, *Syracuse University*
Jill Hamm, *University of North Carolina, Chapel Hill*
John Hattie, *University of Auckland, New Zealand*
Mary Hegarty, *University of California, Santa Barbara*
Jan N. Hughes, *Texas A&M University*
Slava Kalyuga, *University of South Wales, Australia*
Avi Kaplan, *Temple University*
Michael J. Kieffer, *New York University*
Beth Kurtz-Costes, *University of North Carolina, Chapel Hill*
Dan Lapsley, *University of Notre Dame*
Willy Lens, *University of Leuven, Belgium*
Elizabeth A. Linnenbrink-Garcia, *Duke University*
Robert Lorch, *University of Kentucky*
Joseph P. Magliano, *Northern Illinois University*
Andrew Martin, *University of Sydney, Australia*
Andrew J. Mashburn, *Portland State University*
Linda Mason, *Pennsylvania State University*
Richard E. Mayer, *University of California, Santa Barbara*
Charles MacArthur, *University of Delaware*
Catherine McBride-Chang, *The Chinese University of Hong Kong, China*
Nicole M. McNeil, *University of Notre Dame*
Debra K. Meyer, *Elmhurst College*
Keith Millis, *Northern Illinois University*
Alexandre J. S. Morin, *University of Western Sydney, Australia*
Tamera B. Murdock, *University of Missouri, Kansas City*
P. Karen Murphy, *Pennsylvania State University*
Benjamin Nagengast, *Eberhard Karls University of Tübingen*
Mitchell J. Nathan, *University of Wisconsin, Madison*
E. Michael Nussbaum, *University of Nevada, Las Vegas*
Rollanda E. O'Connor, *University of California, Riverside*
Harry O'Neil, *University of Southern California*
Tenaha O'Reilly, *Educational Testing Service*
Philip Parker, *University of Western Sydney, Australia*
Helen Patrick, *Purdue University*

Erika Patall, *University of Texas, Austin*
Reinhard Pekrun, *University of Munich, Germany*
Yaacov Petscher, *Florida State University*
Gary Phye, *Iowa State University*
Keenan Pituch, *University of Texas, Austin*
Jan L. Plass, *New York University*
Patrick Proctor, *Boston College*
Katherine Rawson, *Kent State University*
Robert Renaud, *University of Manitoba, Canada*
Alexander Renkl, *University of Freiburg, Germany*
Catherine Richards-Tutor, *California State University, Long Beach*
Bethany Rittle-Johnson, *Vanderbilt University*
Daniel Robinson, *University of Texas, Austin*
Philip Rodkin, *University of Illinois at Urbana-Champaign*
Christopher A. Sanchez, *Arizona State University*
Katherine Scheiter, *Knowledge Media Research Center, Germany*
Marlene Schommer-Aikins, *Wichita State University*
Gregory Schraw, *University of Nevada, Las Vegas*
Dale Schunk, *University of North Carolina, Greensboro*
Christian D. Schunn, *University of Pittsburgh*
Paula J. Schwanenflugel, *University of Georgia*
Colleen M. Seifert, *University of Michigan*
Timothy Shanahan, *University of Illinois, Chicago*
Gale M. Sinatra, *University of Southern California*
Einar M. Skaalvik, *Norwegian University of Science and Technology, Norway*
John Sweller, *University of New South Wales, Australia*
Keith Thiede, *Boise State University*
Theresa A. Thorkildsen, *University of Illinois, Chicago*
Wendy Troop-Gordon, *North Dakota State University*
Chia-Wen Tsai, *Ming Chuan University-Taiwan*
Timothy Urdan, *Santa Clara University*
Ellen Usher, *University of Kentucky*
Regina Vollmeyer, *University of Frankfurt, Germany*
Jeffrey Walczyk, *Louisiana Technical University*
Charles A. Weaver III, *Baylor University*
Joanna P. Williams, *Columbia University*
Phil Winne, *Simon Fraser University, Canada*
Moshe M. Zeidner, *University of Haifa, Israel*

The main purpose of the *Journal of Educational Psychology*[®] is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit www.apa.org/pubs/journals/subscriptions.aspx

Manuscripts: Submit manuscripts electronically through the Manuscript Submissions Portal found at www.apa.org/pubs/journals/edu according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Art Graesser, Journal of Educational Psychology, 202 Psychology Building University of Memphis, Memphis, TN 38152-3230. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

Copyright and Permission: Those who wish to reuse APA-copyrighted material in a non-APA publication must secure from APA written permission to reproduce a journal article in full or journal text of more than 800 cumulative words or more than 3 tables and/or figures. APA normally grants permission contingent on permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$25 per page. Libraries are permitted to photocopy beyond the limits of the U.S. copyright law: (1) post-1977 articles, provided the per-copy fee in the code for this journal (0022-0663/13/\$12.00) is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923; (2) pre-1978 articles, provided that the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center. For more information along with a permission form, go to www.apa.org/about/contact/copyright/index.aspx

Electronic Access: APA members who subscribe to this journal have automatic access to all issues of the journal in the PsycARTICLES[®] full-text database. See <http://my.apa.org/access.html>.

Reprints: Authors may order reprints of their articles from the printer when they receive proofs.

APA Journal Staff: Susan J. A. Harris, *Senior Director, Journals Program*; John Breithaupt, *Director, Journal Services*; Paige W. Jackson, *Director, Editorial Services*; Stephanie Pollock, *Account Manager*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits*.

Journal of Educational Psychology[®] (ISSN 0022-0663) is published quarterly (February, May, August, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2014 rates follow: *Nonmember Individual*: \$208 Domestic, \$237 Foreign, \$250 Air Mail. *Institutional*: \$751 Domestic, \$800 Foreign, \$815 Air Mail. *APA Member*: \$89. *APA Student Affiliate*: \$62. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Educational Psychology®

www.apa.org/pubs/journals/edu

November 2013

Volume 105
Number 4

Special Issue: Advanced Learning Technologies

Guest Editors: Vincent Aleven, Carole R. Beal, and Arthur C. Graesser

© 2013
American
Psychological
Association

- 929 Introduction to the Special Issue on Advanced Learning Technologies
Vincent Aleven, Carole R. Beal, and Arthur C. Graesser
- 932 Using Adaptive Learning Technologies to Personalize Instruction to
Student Interests: The Impact of Relevant Contexts on Performance and
Learning Outcomes
Candace A. Walkington
- 946 Generalizing Automated Detection of the Robustness of Student Learning
in an Intelligent Tutor for Genetics
Ryan S. J. d. Baker, Albert T. Corbett, and Sujith M. Gowda
- 957 Gender Differences in the Use and Benefit of Advanced Learning
Technologies for Mathematics
*Ivon Arroyo, Winslow Burleson, Minghui Tai, Kasia Muldner,
and Beverly Park Woolf*
- 970 A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on
K–12 Students' Mathematical Learning
Saiying Steenbergen-Hu and Harris Cooper
- 988 Using Student Interactions to Foster Rule–Diagram Mapping During
Problem Solving in an Intelligent Tutoring System
Kirsten R. Butcher and Vincent Aleven
- 1010 Writing Pal: Feasibility of an Intelligent Writing Strategy Tutor in the High
School Classroom
Rod D. Roscoe and Danielle S. McNamara
- 1026 Learning Intercultural Communication Skills With Virtual Humans:
Feedback and Fidelity
H. Chad Lane, Matthew Jensen Hays, Mark G. Core, and Daniel Auerbach
- 1036 Motivation and Performance in a Game-Based Intelligent Tutoring System
G. Tanner Jackson and Danielle S. McNamara
- 1050 The Impact of Individual, Competitive, and Collaborative Mathematics
Game Play on Learning, Performance, and Motivation
*Jan L. Plass, Paul A. O'Keefe, Bruce D. Homer, Jennifer Case,
Elizabeth O. Hayward, Murphy Stein, and Ken Perlin*
- 1067 Guiding Learners Through Technology-Based Instruction: The Effects of
Adaptive Guidance Design and Individual Differences on Learning Over
Time
Adam M. Kanar and Bradford S. Bell
- 1082 A Selective Meta-Analysis on the Relative Incidence of Discrete Affective
States During Learning With Technology
Sidney D'Mello

(Contents continue)

- 1100 Next-Generation Environments for Assessing and Promoting Complex Science Learning
Edys S. Quellmalz, Jodi L. Davenport, Michael J. Timms, George E. DeBoer, Kevin A. Jordan, Chun-Wei Huang, and Barbara C. Buckley
- 1115 My Science Tutor: A Conversational Multimedia Virtual Tutor
Wayne Ward, Ron Cole, Daniel Bolaños, Cindy Buchenroth-Martin, Edward Svirsky, and Tim Weston
- 1126 A Tutoring System That Simulates the Highly Interactive Nature of Human Tutoring
Sandra Katz and Patricia L. Albacete
- 1142 Human and Automated Assessment of Oral Reading Fluency
Daniel Bolaños, Ron A. Cole, Wayne H. Ward, Gerald A. Tindal, Jan Hasbrouck, and Paula J. Schwanenflugel
- 1152 Cognitive Anatomy of Tutor Learning: Lessons Learned With SimStudent
Noboru Matsuda, Evelyn Yarzebinski, Victoria Keiser, Rohan Raizada, William W. Cohen, Gabriel J. Stylianides, and Kenneth R. Koedinger
- 1164 Gendered Socialization With an Embodied Agent: Creating a Social and Affable Mathematics Learning Environment for Middle-Grade Females
Yanghee Kim and Jae Hoon Lim
- 1175 Live Webcam Coaching to Help Early Elementary Classroom Teachers Provide Effective Literacy Instruction for Struggling Readers: The Targeted Reading Intervention
Lynne Vernon-Feagans, Kirsten Kainz, Amy Hedrick, Marnie Ginsberg, and Steve Amendum
- 1188 Using Electronic Portfolios to Foster Literacy and Self-Regulated Learning Skills in Elementary Students
Philip C. Abrami, Vivek Venkatesh, Elizabeth J. Meyer, and C. Anne Wade
- 1210 Universal Design for Learning and Elementary School Science: Exploring the Efficacy, Use, and Perceptions of a Web-Based Science Notebook
Gabrielle Rappolt-Schlichtmann, Samantha G. Daley, Seoin Lim, Scott Lapinski, Kristin H. Robinson, and Mindy Johnson

Other

- iii Acknowledgments
- 1125 Call for Nominations
- 931 Correction to Aleven, Beal, and Graesser (2013)
- 1025 Correction to Hernandez et al. (2013)
- 1226 Instructions to Authors
- 1066 New Editors Appointed, 2015–2020

Introduction to the Special Issue on Advanced Learning Technologies

Vincent Alevén
Carnegie Mellon University

Carole R. Beal
University of Arizona

Arthur C. Graesser
University of Memphis

The 20 articles in this special issue represent a cross-section of interesting, cutting-edge research in advanced learning technologies (ALTs). These advanced technologies are increasingly being used in educational practice and as convenient platforms for rigorous educational research. Although defining ALTs is difficult, ALTs have 3 key elements to varying degrees. First, these technologies are created by designers who have a substantial theoretical and empirical understanding of learners, learning, and the targeted subject matter. Second, these systems provide a high degree of interactivity, reflecting a view of learning as a complex, constructive activity on the part of learners that can be enhanced with detailed, adaptive guidance. Third, the system is capable of assessing learners while they use the system along a range of psychological dimensions. The emphasis in the special issue is not exclusively on the technologies themselves but more fundamentally on the underlying principles of learning, the interactions with the learners, and the impact of the technologies on learning gains. Key challenges are how to develop and use the technologies in ways that are grounded in theory, science, and sensible practice.

Keywords: advanced learning technologies

This special issue presents a group of articles that were submitted in response to a call for papers about advanced learning technologies (ALTs). The articles represent a cross-section of interesting, cutting-edge research in this area. The response to our initial call was overwhelming, testifying to the great research activity this topic has generated. We received over 80 abstracts. We invited 32 full submissions and, eventually, after the *Journal of Educational Psychology*'s stringent peer review process had run its course, we ended up with 20 articles to publish.

Why is a special issue on ALTs both timely and relevant to the readership of the *Journal of Educational Psychology*? There are many good reasons. These advanced technologies are increasingly being used in educational research and practice, as well as in informal learning settings. The advanced technologies are opening doors to new learning experiences. Evidence is accumulating that these technologies can have a substantial positive impact on students' learning outcomes, sometimes even on standardized tests. In addition, new technologies are affecting teacher training and classroom practice.

This collection of articles illustrates how ALTs are convenient platforms for scientific research in addition to addressing applied research questions in rigorous ways. They allow for systematic, consistent, precisely timed administration of instructional interventions. They allow for the recording of data about learning processes at a grain size that can be quite difficult to achieve with other methods, such as hand-coded observations of classroom activities or of students working with technologies. They can record the frequency and detail of learning interactions over prolonged periods of time, in real educational settings, as illustrated in a number of articles in this special issue. As ALTs scale up and become widespread, it is increasingly important to build a strong scientific basis regarding the effectiveness of different technologies. What works well, with what learners, in what contexts?

The emphasis in the special issue is not exclusively on the technologies themselves but more fundamentally on the underlying principles of learning, the interactions with the learners, and the impact of the technologies on learning gains. Simply put, the features of the technologies are grounded in psychological theory and empirically tested in assessments of learning and motivation. The special issue presents a suite of examples of many such technologies and associated theory and evidence. This provides a snapshot of the state of the art in ALTs and an introduction to readers who are not familiar with the field.

It is appropriate to define what we mean by ALTs. There is no authoritative definition of an ALT, but ALTs do have three key elements to varying degrees. First, these technologies are created by designers who have a substantial theoretical and empirical understanding of learners, learning, and the targeted subject matter. Second, these systems provide a high degree of interactivity,

This article was published Online First September 9, 2013.

Vincent Alevén, Human–Computer Interaction Institute, Carnegie Mellon University; Carole R. Beal, School of Information: Science, Technology, and Arts, University of Arizona; Arthur C. Graesser, Department of Psychology, University of Memphis.

Correspondence concerning this article should be addressed to Vincent Alevén, Human–Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: alevén@cs.cmu.edu

reflecting a view of learning as a complex, constructive activity on the part of learners that can be enhanced with detailed, adaptive guidance. Third, the system is capable of assessing learners while they use the system along different psychological dimensions, such as mastery of the targeted domain knowledge, application of learning strategies, and experiences of affective states. On the basis of these assessments, the systems make pedagogical decisions that attempt to adapt to the needs of individual learners. This broad definition encompasses a range of systems and approaches, as exemplified in the current special issue.

It is easy to see how some of the technologies and systems fit our definition of an ALT. For example, intelligent tutoring systems, pedagogical agents, simulations that provide guidance to learners, and the best of educational games (not all of them) fit this definition. For other technologies, it is easy to say that they do not fit the definition. PowerPoint does not fit the definition very well even though it is an advanced tool in many ways. PowerPoint does not adapt to individual learners and rarely is set up to be interactive. SMART Boards are not ALTs in spite of their name. Web cameras in the classroom also do not incorporate adaptation or automatic assessment of student progress. For some systems, it is harder to determine whether or not they should be classified as ALTs. What we can say, however, is that for any given system, the case is stronger to the extent that it exhibits the three elements listed above. Many online courses (e.g., MOOCs, or massive open online courses) contain multimedia content such as video lectures, but the use of multimedia per se does not make them ALTs. A multimedia presentation is not necessarily adaptive and interactive. Online courses can be considered more advanced when they, for example, (a) contain explanations, prompts, and elaborations that are crafted on the basis of theoretical learning principles, a cognitive task analysis, and/or an analysis of student data; (b) decompose student contributions with advanced natural language technologies, affect sensing components, or other intelligent pattern detectors to provide adaptive feedback; and (c) use a sophisticated computational algorithm for deciding what prompts to present to each student, adapting to system-assessed individual differences. Likewise, among the systems that support learners during problem-solving activities, some systems are essentially online worksheets with answers revealed after the learner has solved each problem, whereas others (such as intelligent tutoring systems) can provide highly interactive step-by-step guidance while being sensitive to alternative student strategies and other student variables. Moreover, the pedagogical context matters. What is currently advanced from the point of view of classroom practice may not be advanced from the point of view of researchers. As special issue editors, we have tried to embrace a rather broad set of technologies that are advanced in at least some of these ways.

Roughly half of the articles in this special issue investigate intelligent tutoring systems or systems that integrate intelligent tutoring components with another technology. Examples of the latter are games, simulations, and natural language dialogue technologies that have an added tutoring component. Intelligent tutoring systems are defined as systems that provide detailed guidance (e.g., through hints, feedback, after-action review, or individualized problem selection) as learners work through complex problem scenarios and hone their understanding and problem-solving skills. Of the ALTs featured in the current special issue, intelligent tutoring systems are the ones most commonly found in classrooms.

Nevertheless, the articles in this issue present studies on other learning technologies as well. Some articles present systems capable of understanding natural language and having a dialogue with learners. Others feature systems in which learners interact with computer-based pedagogical agents (talking heads) that provide guidance. In the case of teachable agents, the learner takes on the role of teacher, tapping into the wisdom implicit in the old adage that people never learn so well as when they are teaching. Several articles present research based on educational games, simulations, and virtual learning environments. Finally, the special issue includes articles investigating electronic portfolios and the use of webcam technology that enables teachers to coach struggling readers remotely. The latter technology illustrates our point made above: Although live communication with live video streams over the network is not technologically new, the use of webcams for remote coaching by tutors is an innovative use of standard technology in a classroom setting. Once again, the emphasis is on learning and instruction, not exclusively on technology.

Much of the research in this special issue is experimental research in which innovative technologies are compared against challenging comparison conditions that are less technologically advanced. In some of the articles, the authors test the incremental value and influence of particular technology features on learning. Two studies focused on educational data mining, a new strong trend in research on ALTs. As learning technologies become more prevalent, so do the information-rich data sets collected with these technologies. These data sets can be mined for insight into learning and learning processes. We frequently see that educational data sets are mined in secondary analyses to answer questions that are remarkably different from the ones for which the data were collected. These secondary analyses are often carried out by researchers not involved in the initial data collection. Some researchers have even gone out on a limb by predicting that the main impact of technology on education will be through educational data mining. To some degree, that prediction is borne out in the current set of articles, which provide insight into students' learning processes through log data collected with these technologies. These logs are written automatically by the computer to record interactions between the learners and the computer. Finally, the special issue also contains two meta-analyses that validate the use of novel technologies for tracking assessment of learning and emotions.

Most of the research presented in this issue involves classroom research or secondary analysis of data collected in classrooms. This is interesting in a number of ways. First, results from laboratory experiments do not always survive the transition to classrooms, so it makes sense to work in classrooms to enhance the ecological validity of research studies. Second, the fact that much research took place in real educational settings illustrates a level of maturity of the technologies that are presented. Simply put, they have survived the stringent test of being classroom proof. A class of middle school students working on experimental learning software is a critical audience who will "explore around the edges"! That is, their curiosity leads them to use a learning system in unanticipated ways that are hard to emulate using conventional software testing methods. As a consequence, previously undiscovered flaws in the system are likely to come to light. Third, and particularly important, many of the reported studies with ALTs showed positive learning gains in actual classrooms.

Regarding targeted domains and subject matters, the vast majority of the articles examine challenging K–12 subject matters, namely, reading, writing, math, and science. However, there also are articles in which technologies were used to help learners develop softer skills such as intercultural competence and adeptness at tutoring.

In closing, this collection of articles can be viewed as a harbinger of things to come in the field of educational psychology. No

one doubts that computers will play an increasingly fundamental role in education. The key challenges are how to develop and use the technologies in ways that are grounded in theory, science, and sensible practice. The articles in this special issue are 20 examples providing a glimpse of how this might be accomplished.

Received July 25, 2013

Accepted July 25, 2013 ■

Correction to Aleven, Beal, and Graesser (2013)

In the article “Introduction to the Special Issue on Advanced Learning Technologies” by Vincent Aleven, Carole R. Beal, and Arthur C. Graesser (*Journal of Educational Psychology*, Advance online publication. September 9, 2013. doi: 10.1037/a0034155), the name of author Arthur C. Graesser was published with an incorrect middle initial. All versions of this article have been corrected.

DOI: 10.1037/a0034715

Using Adaptive Learning Technologies to Personalize Instruction to Student Interests: The Impact of Relevant Contexts on Performance and Learning Outcomes

Candace A. Walkington
Southern Methodist University

Adaptive learning technologies are emerging in educational settings as a means to customize instruction to learners' background, experiences, and prior knowledge. Here, a technology-based personalization intervention within an intelligent tutoring system (ITS) for secondary mathematics was used to adapt instruction to students' personal interests. We conducted a learning experiment where 145 ninth-grade Algebra I students were randomly assigned to 2 conditions in the Cognitive Tutor Algebra ITS. For 1 instructional unit, half of the students received normal algebra story problems, and half received matched problems personalized to their out-of-school interests in areas such as sports, music, and movies. Results showed that students in the personalization condition solved problems faster and more accurately within the modified unit. The impact of personalization was most pronounced for 1 skill in particular—writing symbolic equations from story scenarios—and for 1 group of students in particular—students who were struggling to learn within the tutoring environment. Once the treatment had been removed, students who had received personalization continued to write symbolic equations for normal story problems with increasingly complex structures more accurately and with greater efficiency. Thus, we provide evidence that interest-based interventions can promote robust learning outcomes—such as transfer and accelerated future learning—in secondary mathematics. These interest-based connections may allow for abstract ideas to become perceptually grounded in students' experiences such that they become easier to grasp. Adaptive learning technologies that utilize interest may be a powerful way to support learners in gaining fluency with abstract representational systems.

Keywords: intelligent tutoring system, personalization, individual interest, topic interest, algebra

The computer is the Proteus of machines. Its essence is its universality, its power to simulate. Because it can take on a thousand forms and can serve a thousand functions, it can appeal to a thousand tastes. (Papert, 1980, p. xxi)

Advanced learning technologies are emerging in educational settings as a powerful means to adapt instruction to learners' backgrounds, goals, preferences, and prior knowledge (Papert, 1980, 1993). Such technologies have the potential to transform the very nature of education by allowing for a level of customization that can fundamentally change the relationship between the learner and the content to be learned (Collins & Halverson, 2009). In

particular, learning technology innovations allow for instruction to be personalized to users' actions and interests, to provide assistance when needed and present instruction that is understandable, engaging, and situated in relevant and meaningful contexts. One well-known example of this type of personalized learning is intelligent tutoring systems (ITSs)—technology environments that utilize artificial intelligence to adapt instruction to learner knowledge states (Koedinger & Corbett, 2006). As Papert (1980) predicted, the computer has truly become the "Proteus of machines," a shape-shifter able to adapt to the infinite variations in the background and preferences of its users.

The rise of such adaptive learning technologies is timely, given the pressing issues with motivation that face schools today (Hidi & Harackiewicz, 2000). One important principle for adaptive learning environments is that instruction may be effective when presented in the context of learners' *interests*—their predispositions to engage with particular topics, ideas, or activities (Hidi & Renninger, 2006). Research shows that presenting instruction in the context of learner's interests is effective, impacting persistence, attention, and engagement (e.g., Ainley, Hidi, & Berndorff, 2002; Flowerday, Schraw, & Stevens, 2004; Hidi, 1990, 2001). However, with the exception of a few studies in reading (e.g., Heilman, Collins, Eskenazi, Juffs, & Wilson, 2010) and basic mathematics (e.g., Anand & Ross, 1987; Cordova & Lepper, 1996), there has been little research on bringing student interests into adaptive technology-based learning environments.

This article was published Online First September 9, 2013.

This work was supported in part by the Pittsburgh Science of Learning Center, which is funded by National Science Foundation Award SBE-0354420. This work was carried out in partnership with Carnegie Learning. This work was also supported in part by the Postdoctoral Training Program in Mathematical Thinking, Learning, and Instruction, funded by U.S. Department of Education Institute of Education Sciences Award R305B100007. The contributions of Milan Sherman, Anthony Petrosino, Jim Greeno, Mitchell Nathan, and Ken Koedinger to this work are greatly appreciated.

Correspondence concerning this article should be addressed to Candace A. Walkington, Department of Teaching and Learning, Southern Methodist University, P.O. Box 740455, Dallas, TX 75275. E-mail: cwalkington@smu.edu

One domain that may benefit from interest-based interventions is high school algebra. Algebra has been framed as a gatekeeper to higher level mathematics, with significant implications for students' economic futures (Kaput, 2000; Moses & Cobb, 2001). Well-known issues with motivation and interest occur during adolescence and secondary mathematics courses (Fredricks & Eccles, 2002; Frenzel, Goetz, Pekrun, & Watt, 2010; McCoy, 2005; Mitchell, 1993). In algebra, students make an important transition from working with known quantities to using symbols to represent unknown quantities, learning skills like writing, manipulating, and solving algebraic expressions (Common Core State Standards Initiative, 2010). Interest interventions may be designed to provide a type of perceptual *grounding* (Goldstone & Son, 2005) for abstract systems of representation, making them more situated and understandable. Grounding is accomplished when abstract ideas are related to concrete objects or events that learners are familiar with, like their interests. Goldstone and Son (2005) found that initial presentation of scientific principles in concrete, grounded form improves transfer when this concreteness is faded over time.

When considering how concrete, interest-based representations of ideas may benefit students, it is important to account for the desired learning outcomes. An intervention designed to make concepts easier and more approachable will not necessarily enhance student learning. This distinction is captured by research on *desirable difficulties*, which has shown that modifications that make a task more difficult during training, like decreasing feedback, can actually enhance learning measures like retention (Schmidt & Bjork, 1992). To distinguish between immediate performance during training and students' long-term learning, Koedinger, Corbett, and Perfetti (2012) identified three types of *robust learning*: learning that is retained over time (long-term retention), learning that can be applied in new situations (transfer), and learning that can form the basis for new concepts (accelerated future learning). An important goal of research on robust learning is to explore instructional principles associated with improved outcomes while also determining how these principles are impacted by student-level and knowledge-level characteristics in different content areas.

Interest-based interventions designed to promote grounding have the potential to improve robust learning because they not only make tasks easier for students to conceptualize but may promote worthwhile and robust connections between students' prior knowledge and abstract systems of representation. In particular, embedding instruction in students' interests may facilitate connections between students' *situation models* of the actions, events, and relationships in the mathematical scenario they are confronting and the associated *problem models* containing formal notation (Kintsch, 1986; Nathan, Kintsch, & Young, 1992). These connections may allow students to use abstractions effectively and meaningfully, while they remain relatively portable to a variety of situations.

Here, we seek to contribute to the theoretical and empirical bases of interest-based interventions by expanding this research to a new, more abstract domain and by examining how a robust psychological principle of learning, topic interest, can be integrated into adaptive technologies to promote learning. We explore the instructional principle of *context personalization* (or now, personalization for short), which is a type of interest-based intervention where instructional contexts are matched to students' out-of-school interests (Anand & Ross, 1987; Cordova & Lepper,

1996). This work takes place within Cognitive Tutor Algebra (CTA), an ITS that already contains powerful capabilities to adapt instruction to student knowledge. We look at how the resources of an ITS can be further leveraged to connect instruction to interests and discuss the outcomes of this intervention for robust learning.

Theoretical Framework

Individual, Situational, and Topic Interest

Context personalization is an instructional intervention that is hypothesized to mediate learning outcomes by eliciting *interest* (Heilman, Collins, Eskenazi, Juffs, & Wilson, 2010; Reber, Hetland, Chen, Norman, & Kobbeltvedt, 2009; Renninger, Ewen, & Lasher, 2002). Hidi and Renninger (2006) defined interest as "the psychological state of engaging or the predisposition to reengage with particular classes of objects, events, or ideas over time" (p. 112) and accentuated that it has cognitive and affective components. Interest unfolds as learners interact with their environment (Renninger & Hidi, 2011). The personalization intervention here is designed to elicit *topic interest*—interest triggered when learners are presented with a specific topic or theme. Topic interest is dependent on characteristics of both the learner and the environment and has aspects of both *individual* and *situational* interest (Ainley et al., 2002; Hidi, 2001).

Individual interest is the relatively stable and enduring preferences held by a learner toward specific activities, objects, or events. Individual interest is composed of both *stored value*, the learner's feelings toward the activity, and *stored knowledge*, the learner's understanding of the structure and discourse of the activity (Renninger et al., 2002). Situational interest, on the other hand, is an attention-focusing and affective reaction to particular stimuli or characteristics of the environment, such as coherence, salience, personal relevance, or incongruity (Hidi & Renninger, 2006). Situational interest can be triggered as an intervention grabs students' attention and maintained as students engage with the material (Linnenbrink-Garcia et al., 2010). Our personalization intervention is both an environmental modification intended to trigger and maintain situational interest and a way of leveraging individual interests in out-of-school topics. Learners may need supports like personalization to initially connect to the content, but such connections may need to be meaningful to maintain interest. The environment and the learner's goals and characteristics are critical to the development of interest (Renninger & Su, 2012).

Interest and Prior Knowledge

Instructional modifications that involve individual interest necessarily involve prior knowledge—learners are likely to have high prior knowledge about their interests (Renninger et al., 2002). The effect of interventions that leverage both knowledge and value-related components of interest may be especially powerful because the interest-based triggers are tied explicitly to the content to be learned (Mitchell, 1993; Reber et al., 2009). This is the approach of the intervention used here, and it is contrasted with interest-based triggers that are not relevant to the learning task, such as decorative pictures or the insertion of incidental elements like student names (e.g., Cordova & Lepper, 1996). This distinction is especially important in algebra, considering that meaningfulness of

the mathematical content to secondary students' lives is an important component of interest and that many students find mathematics courses to be without such meaning (Mitchell, 1993).

Interest and Instructional Outcomes

Interest can be developed through carefully designed learning environments that allow students to connect to the content they are learning (Renninger & Hidi, 2011). The activation of interest is associated with improved learning (Ainley et al., 2002; Ainley, Hillman, & Hidi, 2002; Harackiewicz, Durick, Barron, Linnenbrink, & Tauer, 2008; Schiefele, 1990, 1991), as well as with increased attention (McDaniel, Waddill, Finstad, & Bourg, 2000; Renninger & Wozinak, 1985), persistence (Ainley et al., 2002), engagement (Flowerday et al., 2004), reported task involvement and perceived competence (Durik & Harackiewicz, 2007), reported utility value (Hulleman, Godes, Hendricks, & Harackiewicz, 2010), and motivational variables like self-efficacy, self-regulation, and achievement goals (Harackiewicz et al., 2008; Hidi & Ainley, 2008; Sansone, Fraughton, Zachary, Butner, & Heiner, 2011). Interest-based interventions can promote academic achievement and interest in future courses and careers (Cordova & Lepper, 1996; Harackiewicz et al., 2012; Hulleman & Harackiewicz, 2009).

To investigate the path from activated interest to increased learning, researchers have used response time measures to examine attention and persistence. Ainley et al. (2002) found that topic interest was associated with students continuing to read a text rather than stopping reading, which in turn affected learning outcomes. However, other work has found faster response times when students have interest in the topic, suggesting that interest facilitates automatic allocation of attention and frees up cognitive resources (Hidi, 1990; McDaniel et al., 2000). Such time measures have not been well examined in the domain of mathematics, so from previous literature, it is unclear whether an interest-based intervention would increase problem-solving time by enhancing persistence or reduce problem-solving time by facilitating attentional allocation. Koedinger et al. (2012) discussed measuring *learning efficiency* in interventions—the idea that since instructional time is so valuable in classrooms, completing activities in less time without reducing learning is an important outcome. They observed that “too many theoretical analyses and experimental studies do not address the time costs of instructional methods” (Koedinger et al., 2012, p. 34).

Working within an ITS system, there are additional behavioral measures that may be important when considering the relationship among interest, attention, and persistence. Renninger and Su (2012) described how students' connection to or interest in the content can impact how they make use of available supports. ITS research has shown that learners can engage in “gaming-the-system” behaviors where they take advantage of the tutor's hints and feedback. Learners may quickly click through all of the hints available for a problem, until they reach the bottom-out hint giving the answer, or they may enter different answers quickly and repeatedly, looking for the response the tutor will accept. Gaming behaviors have been found to be negatively correlated with learning (Baker, Corbett, Koedinger, & Wagner, 2004). Here, an examination of both learning efficiency and gaming behaviors is provided in order to explore how interest may interact with attention and persistence.

Interest and Mathematics Learning

An interest-based intervention may be especially effective for mathematics learning. Koedinger and Nathan (2004) found that algebra story problems and verbal word equations were easier for students to solve than matched symbolic equations. One explanation for this phenomenon is that verbal contexts allow for abstract mathematical ideas to become grounded in concrete experiences, fostering important connections to prior knowledge (Wilensky, 1991). Kintsch's (1986) model of text comprehension can be used to hypothesize how interest and prior knowledge may impact learning from story problems. In this model, learners create a *textbase* of the propositions supplied directly in the text, as well as a *situation model* that integrates the content of the text with prior knowledge. Schiefele (1999) reported that a reader's level of interest in a text's topic is more highly correlated with deep-level processing measures (e.g., comprehension) than surface-level processing measures (e.g., recall). Similarly, McDaniel et al. (2000) found that interest increased the learner's focus on organizing the text using structural linkages (perhaps to construct situation models), instead of focusing on simply extracting the proposition-specific content. Thus, activated interest may be associated with more meaningful, detailed, and accurate situation models.

Kintsch (1986) accentuated that when learners are familiar with the situation being described in a mathematics word problem, they are more likely to correctly formulate a solution. Nathan et al. (1992) found that when students' construction of situation models was supported through animations of the actions and relationships in the story, students were better able to write equations from story problems. Activating interest may be an important method for supporting learners in successfully coordinating situation and problem models, providing a means of grounding abstract ideas. In an investigation of children solving personalized arithmetic problems, Renninger et al. (2002) found that personalized scenarios allowed students to form potentially powerful connections between the context of the story problem and the underlying mathematics content. They discussed how problems with interest-based contexts can allow some students to focus on the meaning of the problems being posed and away from keywords. Similarly, when investigating the insertion of interest-based contexts into reading passages, they found that these contexts allowed some students to focus on extracting meaning from the text and provided a scaffold for decoding and recall. Interest-based contexts may have provided a means to ground the deeper structure of these tasks in students' prior knowledge.

Interest may promote grounding by allowing learners to form meaningful connections between their qualitative understanding of story scenarios and their attempts to mathematically model these situations. This could be contrasted with *direct translation* or keyword-type approaches where learners map directly from the propositions in the word problem text to a series of computations (Hegarty, Mayer, & Monk, 1995). The strength of interest-based interventions in mathematics may be related to the support they provide for both the construction of accurate and meaningful situation models and the successful coordination of situation and problem models during problem solving.

Literature Review

Studies of Personalization in Mathematics

There have been a number of previous studies on personalization in elementary mathematics, which have had mixed results. Two studies personalized incidental aspects (e.g., inserting students' name and favorite food) of computer environments that provided instruction on arithmetic and found that students who received personalization outperformed control groups on measures of learning (Anand & Ross, 1987; Cordova & Lepper, 1996). Two other studies found posttest differences for personalization using arithmetic story problems (Chen & Liu, 2007; López & Sullivan, 1992), while an additional study found that personalized arithmetic problems were easier for students (Davis-Dorsey, Ross, & Morrison, 1991). However, several studies found no effect for personalization (E. Bates & Wiest, 2004; Cakir & Simsek, 2010; Ku & Sullivan, 2000). All of these studies involved elementary school content—usually arithmetic problems. The present study is unique in that it extends this work into algebra and secondary mathematics. As such, we next briefly review the literature on the development of algebraic reasoning.

The Development of Algebraic Reasoning

Key to the development of algebraic reasoning is working with and using variables to represent unknown quantities. Symbolization and symbol manipulation tasks are challenging for students to learn (Filloy & Rojano, 1989; Herscovics & Linchevski, 1994; Koedinger & Nathan, 2004; Stacey & MacGregor, 1999; Walkington, Sherman, & Petrosino, 2012). When students write symbolic equations of functional relationships, they often view these equations as a sequence of calculations rather than a statement about equality (Breidenbach, Dubinsky, Hawks, & Nichols, 1992; Clement, 1982; Humberstone & Reeve, 2008; Stacey & MacGregor, 1999). Students erroneously assign shifting or multiple values to unknowns and may allow one variable to stand for two different quantities (Stacey & Macgregor, 1999). As such, students have difficulty conceptualizing the idea of operating directly on an unknown quantity (Filloy & Rojano, 1989; Herscovics & Linchevski, 1994; Stacey & MacGregor, 1999).

The present study focuses on writing symbolic expressions from story scenarios, which is a particularly difficult concept (Bardini, Pierce, & Stacey, 2004; Heffernan & Koedinger, 1997; Koedinger & McLaughlin, 2010; Nathan et al., 1992; Swafford & Langrall, 2000; Walkington et al., 2012). Although students may be able to describe the relationships in a story problem verbally, using standard algebraic notation is more challenging (Bardini et al., 2004; Swafford & Langrall, 2000). Students must learn to negotiate the mathematical “grammar of such expressions” and can have difficulty combining or composing different parts of expressions (Koedinger & McLaughlin, 2010, p. 471; see also Heffernan & Koedinger, 1997). Students may not see the utility of writing an equation from a story scenario and can have a tendency to solve for concrete, specific cases without formulating a general equation (Stacey & MacGregor, 1999; Swafford & Langrall, 2000).

Despite the difficulties learners encounter with algebraic symbolization, research has shown that using relevant contexts based in students' experiences is an effective method for introducing

tools of abstraction (Bardini et al., 2004; Carraher, Schliemann, Brizuela, & Earnest, 2006; Chazan, 1999; Lampert, 2001; Moses & Cobb, 2001). These connections may support students in making the difficult transition to using symbols, especially when contexts are linked to students' interests and experiences through personalization. As described previously, personalized contexts may activate interest, allowing for greater focus of attention, engagement, and persistence in the difficult task of algebraic symbolization. Interest-based scenarios may also ground abstract ideas in concrete experiences and prior knowledge. Thus, personalization has the potential to support the development of algebraic reasoning.

Personalization in Algebra

In prior work, we presented 24 ninth-grade Algebra I students with a set of algebra story problems that contained both normal problems and problems personalized to student interests (Walkington, Petrosino, & Sherman, in press). Results showed that when solving personalized problems, students more often used informal strategies that closely mirrored the action of the story (15% of responses for normal problems, 42% of responses for personalized problems). Students also made fewer conceptual mistakes when mathematically formulating the relationships described in a personalized story (24% omitted intercept for normal problems, 13% omitted intercept for personalized problems). Regression models indicated that personalization had a significant and positive impact on performance for struggling students who performed poorly on their problem set ($p < .01$) and for problems that incorporated a particularly difficult linear function, such as “ $y = -0.23x + 7.87$ ” ($p < .05$). There were no significant differences related to student demographics. For more information on this study's methodology and results, see Walkington et al. (in press).

Based on this work, we hypothesized that the dependence of the effect of personalization on student ability and problem difficulty may be the reason why personalization studies have shown mixed results. Personalization might be particularly effective in the context of an ITS, where problems that incorporate different skills are selected at the appropriate difficulty level based on a cognitive model of student performance. Here, we conducted an experiment in the CTA environment where students were randomly assigned to receive story problems personalized to their interests or standard story problems for one unit. Personalization was accomplished by administering an interests survey to both groups and having the experimental group receive problem variations corresponding to their areas of interest in topics like sports, music, and movies. We examined students' immediate performance on the personalized and normal problems, as well as their robust learning from the intervention, in terms of accelerated future learning and transfer to novel problem formats. We also looked at how receiving personalized problems impacted learning efficiency measures.

Research Questions

We pursued three research questions.

R.1: How does context personalization impact performance measures while the intervention is in place? In particular, how is this effect moderated by (a) the ability of the learner and (b) the difficulty of the task? Based on prior work, we hypothesized that personalization would enhance performance and that this

effect should be strongest when students are struggling to learn difficult concepts.

R.2: How does context personalization impact response time measures while the intervention is in place? Although research on the impact of interest on problem-solving duration is mixed, we hypothesized that personalization would reduce time spent solving problems. Personalization may impact other time measures, such as the tendency to game the system by taking advantage of ITS feedback.

R.3: How does context personalization impact measures of robust learning once the intervention is removed? Based on the literature on interest, grounding, and personalization, we hypothesized that personalized contexts would help students understand the underlying concepts, allowing for transfer to and accelerated future learning in future topics related to linear functions in the curriculum.

Method

Participants

The high school at which the study took place was in a rural area outside a large city in the northeastern United States. The school was 96% Caucasian and 51% male/49% female, and 18% of students were eligible for free/reduced lunch. There were three Algebra I teachers at the school who taught nine different Algebra I classes. Algebra I students at the school were usually classified as ninth graders; however, a few 10th and 11th graders were enrolled in Algebra I as repeaters. The experiment impacted Unit 6 of the CTA software. The experiment was “in sequence,” meaning that students reached Unit 6 in CTA at their own pace as they worked in the tutoring environment 2 days each week. There were 195 Algebra I students with CTA accounts, and of these, 145 (73 control, 72 experimental) were included in the study. The excluded students either had completed Unit 6 prior to the study’s start in October ($n = 36$) or did not reach Unit 6 by the end of the school year ($n = 14$). Since curriculum progress in CTA can be considered a proxy for student achievement,¹ the data analyses for Unit 6 omit a group of the top-performing students and a few of the very weak students in these classes. The CTA data set that was used for this study was deidentified, meaning that it contained no information on student background variables like gender and ethnicity. This is a limitation of this type of data; however, we have no reason to believe that the students included in the study did not generally reflect the demographic characteristics of the school, especially given the relative uniformity of the school’s demographic makeup.

Learning Environment

The study took place within the CTA software environment. CTA is an ITS for Algebra I that uses model-tracing approaches to individualize problem selection and knowledge-tracing approaches to individualize hints and feedback (Koedinger & Aleven, 2007). CTA focuses on mathematical functions, and students must negotiate different representational formats (equations, tables, graphs) using computational tools (equation-solving tool, spreadsheet). We modified one unit in CTA—Unit 6, Linear Models and Independent Variables. This unit contains story problems that involve linear

functions (see Figure 1) where students write symbolic expressions and fill in tables with functional values. There were 27 story scenarios in Unit 6, and each could contain several slightly different sets of numbers. As is typical for an ITS, the number of problems each student received was dependent on the student’s performance. On average, each student received 24.39 different problems ($SD = 9.46$ problems) and spent a total of 3 hr and 39 min in Unit 6 ($SD = 2$ hr and 25 min) over 6.38 different days ($SD = 3.76$ days).

In order to assess robust learning through measures of transfer and accelerated future learning, students’ performance in Unit 10 of CTA, Linear Models and Four Quadrant Graphs, was also examined. Unit 10 was the next unit in CTA that covered similar content to Unit 6, containing story scenarios on linear functions. The functions in Unit 10 were more complex than in Unit 6, as they were more likely to include fractions and negative terms. Thus performance in Unit 10 could measure transfer of concepts learned in Unit 6 to more complex linear functions, and time measures in Unit 10 could measure accelerated future learning. All story problems in Unit 10 were nonpersonalized, meaning that students in the experimental group would need to transfer the skills they learned from solving personalized problems in Unit 6 to more complex normal problems in Unit 10. It could be argued that the experimental condition was at a disadvantage when solving normal problems in Unit 10 since the control condition had received normal problems all along during Unit 6. Students typically reached Unit 10 one or two months after Unit 6 ($M = 1.01$ months, $SD = 0.77$ months). Out of the 145 students who were randomly assigned to control or experimental conditions within Unit 6, 122 made it to Unit 10 before the school year ended. Thus, the analyses from Unit 10 likely omitted additional low-performing students ($N = 23$).

Knowledge Components in CTA

The CTA environment adapts instruction by tracking students’ learning of different *knowledge components* (KCs; Koedinger & Aleven, 2007) or mathematical concepts. While the curricular goals of CTA come from state standards, the KCs are at a finer grained level. They are initially developed by the curriculum designers and refined as student data sets are analyzed. Due to the large number of KCs assessed in each unit, for the analyses KCs were grouped into three categories—easy, medium, and hard. This grouping was based on students’ actual performance on different KCs in Units 6 and 10 (shown in Tables 1–2). The classifications of easy/medium/hard corresponded approximately to performance quartiles of 25%–50%, 50%–75%, and 75%–100% correct. In Unit 6, two KCs were classified as hard (writing symbolic expressions that involved a positive slope and intercept or a negative slope and intercept), and in Unit 10, expression writing was also classified as hard.

Materials and Tasks

Before entering Unit 6, students completed an interests survey within CTA where they rated their level of interest (“How much do

¹ Curriculum progress is considered by the developers of CTA to be the best measure of student knowledge and achievement within the Cognitive Tutor environment (S. Ritter, personal communication, 2008). Curriculum progress is also a good measure of achievement because it is tied to students’ Algebra I course grades.

The screenshot shows the Cognitive Tutor Algebra software interface. The top menu bar includes 'File', 'Tutor', 'Go To', 'View', and 'Help'. The main window is titled '8 - Linear Models and Independent Variables' and '1 - Finding Independent Variables with Positive Rates of Change'. The 'Scenario' pane on the left contains a word problem about a raise at PAT-E-OH Furniture Inc. and four questions. The 'Worksheet' pane on the right has a table for 'Quantity Name', 'Unit', 'Expression', and 'Question 1' through 'Question 4'. An 'Answer Key' table is overlaid on the right side of the worksheet.

Scenario

You have just been promoted to assistant manager at PAT-E-OH Furniture Inc. and have received a raise to \$10.50 per hour.

- How much would you be paid if you worked five hours?
- How much would you be paid if you worked 10 and 1/2 hours? If you have not already done so, please fill in the expression row with an algebraic expression for the total pay. Then use the expression and the Solver to answer questions 3 and 4 below.
- How many hours must you work to make five hundred fifty dollars?
- In order to make \$2,200.00, how many hours must you work?

To write the expression, define a variable for the time worked and use this variable to write a rule for your total pay.

Worksheet

Quantity Name	Unit	Expression	Question 1	Question 2	Question 3	Question 4

Answer Key:

Quantity Name	the time worked	the money earned
Unit	hour	dollar
Expression	X	10.5X
Question 1	5	52.5
Question 2	10.5	110.25
Question 3	52.381	550
Question 4	209.5238	2200

Figure 1. Screenshot of a question displayed in Unit 6 of the Cognitive Tutor Algebra software. The Answer Key table has been superimposed over the screenshot to show the correct answers to each question.

you like [topic]?") in nine different areas (sports, music, movies, TV, games, art, computers, food, and stores) on a 4-point scale (1 = *It's boring*, 2 = *It's okay*, 3 = *I like it*, 4 = *It's my favorite thing*). This assessment has been used in our prior work in conjunction with student interviews and is intended to measure level of interest in or liking of (Renninger & Hidi, 2011) topic areas. Four variations on each of the 27 original story problems in Unit 6 were written to correspond to different topic interests. The variations had the same mathematical structure (i.e., linear function) as the original problem, but they had different cover stories relating to different interest categories (see Table 3). Only one set of numbers is shown in Table 3 for simplicity; however, in this and other actual problems, CTA used variations with slightly different number sets to prevent cheating. This was controlled for in the regression models.

Personalized variations were written based on open-ended interest surveys conducted in area high schools ($N = 50$) and additional surveys ($N = 22$) and interviews ($N = 29$) conducted in prior work. Problems were constructed by taking specific objects, events, or ideas that were mentioned in relation to a topic during the surveys or interviews and writing algebra problems that corresponded to these discussions. Two master teachers of algebra reviewed the personalized scenarios for their understandability and relevance to high school students, and modifications were made accordingly. The personalized problems were thus handcrafted for this study. Problem creation can be a bottleneck for the development of adaptive interventions. However, large banks of personalized problems associated with a full mathematics curriculum were recently created for the new ITS *MATHia* (Carnegie Learning, 2011), which personalizes problems to student interests such as sports and music and is already in use in over 200 schools. This suggests that developers may see the potential of such an investment.

Procedures

Before entering Unit 6, all participants were given the interests survey. Participants were then randomly assigned to control or treatment groups. The control condition received the normal algebra story problems already used within Unit 6, while the treatment condition received one of four possible interest-related variations of each problem, based on responses to the interest survey. The com-

Table 1
Difficulty of Knowledge Components Assessed Within Unit 6 in Terms of Student Accuracy

Knowledge component	Percent correct	Knowledge component classification
No knowledge component	88.78	Easy
Identifying independent and dependent units	85.47	Easy
Entering a given	85.31	Easy
Solving for x , slope only	75.96	Medium
Using difficult numbers	74.96	Medium
Using small numbers	74.76	Medium
Solving for x , negative slope with intercept	69.70	Medium
Solving for x , positive slope with intercept	68.74	Medium
Write expression, slope only	64.79	Medium
Write expression, negative slope with intercept	50.86	Hard
Write expression, positive slope with intercept	44.45	Hard

Note. No knowledge component cells included labeling independent and dependent quantities.

Table 2
Difficulty of Knowledge Components Assessed Within Unit 10 in Terms of Student Accuracy

Knowledge component	Percent correct	Knowledge component classification
No knowledge component	92.14	Easy
Identifying units	84.72	Easy
Creating graphs of linear functions	82.33	Easy
Entering a given	79.40	Easy
Using simple numbers	71.13	Medium
Using large numbers	70.08	Medium
Using small numbers	69.80	Medium
Using difficult numbers	69.49	Medium
Solving for x , fraction/negative slope	60.73	Medium
Solving for x , positive integer slope	51.47	Medium
Write expression, any form	42.33	Hard

Note. No knowledge component cells included labeling independent and dependent quantities.

puter would choose the variation that had the highest rated level of interest on the interests survey. Personalization was only in place for Unit 6, and performance and learning measures were collected for Units 6 and 10.

Data Analysis

CTA collects detailed logs of students' interactions with the tutoring environment, including how they answered each problem, hints and feedback received, and response times. Student logs for Unit 6 and Unit 10 of the software were analyzed with multilevel models (Snijders & Bosker, 1999) using the R software package with the lmer function (D. Bates & Maechler, 2010). This technique was used because it allows for logistic modeling of dichotomous outcomes (correct/incorrect response) without the requirement that the data be balanced, fully nested, or fully crossed. Here, because of the adaptive nature of the CTA program and the personalization, not all students would receive the same problems or the same number of problems.

The Level 1 unit of analyses was repeated observations of each student solving one part of one problem (i.e., filling in one cell in Figure 1) involving one KC. Random effects included students nested within teachers, which story problem and number set the student was working on, and the item or linear function (i.e., $y = -2.5x - 35$ in Table 1) underlying that story problem.² Fixed effects included which condition the student had been assigned to in Unit 6 (experimental or control), the difficulty of the KC being tracked in the problem part (easy, medium, hard), and the interaction of condition and KC. The dependent measure was either whether the student got the problem part correct or incorrect on the first attempt (a logistic regression model) or the number of seconds the student took to enter an answer to the problem part (a linear regression model).

Analyses of *learning curves* (Mathan & Koedinger, 2005) for KCs were also conducted. Learning curves show average levels of performance on KCs dependent on the number of opportunities students have had to practice, which is the number of times they have been given a problem part that involves the KC. An analysis

of whether students in the experimental condition were able to learn KCs with less practice was conducted by adding the number of practice opportunities and the interaction of condition with practice opportunities to the model. For an overview of how these equations are modeled and their assumptions, see Snijders and Bosker (1999).

Results

This section examines how students performed in Units 6 and 10 in terms of correct responses. We look at performance on different KCs and learning curves of how experimental and control students mastered concepts over time. We examine how these results varied for students struggling with algebra. We then look at response time measures, including gaming the system. All of these measures are given for Unit 6 (while the intervention was in place) and Unit 10 (after the intervention was removed).

The personalization intervention in Unit 6 improved student accuracy and decreased response times for hard KCs involving writing expressions. In Unit 10, after the intervention had been removed for four units, students in the experimental group were still more accurate and faster at writing more complex expressions, suggesting transfer and accelerated future learning. Within Unit 6, students in the personalization condition learned hard KCs in fewer attempts, and the impact of personalization was significantly greater for students identified as struggling with algebra. We now describe each of these results in more detail in the following sections.

Performance Measures in Units 6 and 10

Correct responses. Table 4 shows that students in the personalization condition performed significantly better than the control group on hard KCs within Unit 6 (odds = 1.53, $p < .001$). Transforming odds into probabilities, the control group's accuracy when writing expressions was 43.20% ($0.76/[1 + 0.76]$), while the experimental group had a 53.66% success rate ($0.76 \times 1.52/[1 + 0.76 \times 1.52]$). Personalization also significantly improved performance on easy KCs in Unit 6 (odds = $1.52 \times 0.94 = 1.43$, $p < .001$). Personalization may support fluidity on less mathematically relevant parts of problems, like entering givens and labeling quantities. Personalization had a nonsignificant but directionally positive effect on performance for medium KCs (odds = $1.52 \times 0.77 = 1.17$, $p = .06$).

Unit 10 performance is shown in Table 5. An additional fixed effect was added for the number of opportunities each student had to solve hard KCs in Unit 6, to control for differential exposure between conditions.³ However, results were similar with or without this term. As shown in Table 5, four units later, with the

² The item variable was only necessary for Unit 6, since within Unit 6, the same item could have different cover stories depending on whether the student was in the experimental or control group and which interests were selected.

³ The median number of hard KC problem parts presented to students in the control group in Unit 6 was 22 ($M = 20.53$), compared to 19.5 ($M = 22.97$) for the experimental group. Being given more opportunities to master a KC, rather than being advanced to the next section, usually evidences weaker performance.

Table 3

Example of Story Problem From Unit 6 Received by Students in the Control Group, Followed by Four Interest-Based Variations of This Problem That CTA Chose From for Students in the Experimental Group

Interest	Problem text
Normal problem (control group)	An experimental liquid (LOT#XLHS-240) is being tested to determine its behavior under extremely low temperatures. Its current temperature is -35 degrees Celsius and is slowly being lowered by two and one-half degrees per hour. What will the temperature of the liquid be ten hours from now? What will the temperature of the liquid be tomorrow at this time? When will the temperature drop to one hundred degrees below zero Celsius? Assuming that the temperature has been dropping at the same rate, when was the temperature zero degrees Celsius?
Food	A new soda at McDonald's is being tested to determine its behavior under extremely low temperatures. Its current temperature is -35 degrees Fahrenheit and is slowly being lowered by two and one-half degrees per hour.
Sports	A new sports drink is being tested to determine its behavior under extremely low temperatures. Its current temperature is -35 degrees Fahrenheit and is slowly being lowered by two and one-half degrees per hour.
Stores	The Dippin' Dots store at the mall uses extremely low temperatures to freeze its ice cream into tiny balls. Right now, the temperature of a batch of chocolate Dippin' Dots ice cream is -35 degrees Fahrenheit and is slowly being lowered by two and one-half degrees per hour.
Movies	The Dippin' Dots stand at the movie theater uses extremely low temperatures to freeze its ice cream into tiny balls. Right now, the current temperature of a batch of chocolate Dippin' Dots ice cream is -35 degrees Fahrenheit and is slowly being lowered by two and one-half degrees per hour.

Note. In this and some other problems, the units were changed as part of the personalization (e.g., American high school students have everyday experience considering temperature in degrees Fahrenheit, not degrees Celsius). CTA = Cognitive Tutor Algebra.

experimental treatment long removed,⁴ students who had received personalization in Unit 6 were still significantly better at algebraic expression writing in Unit 10 (odds = 1.30, $p = .0097$). The knowledge students gained from receiving personalization in Unit 6 may have transferred to more difficult tasks involving writing expressions with negative or fraction slopes from normal story problems. There were no significant differences for easy ($p = .48$) or medium ($p = .42$) KCs in Unit 10.

Learning curves. Learning curves show average performance levels for a KC dependent on the number of opportunities students have had to practice that KC. Models that use learning curves are important because they control for number of practice opportunities, taking into account the possible effects of differential exposure to KCs. The learning curves for students solving hard KCs in Unit 6 are shown in Figure 2. The lack of a clear upward trend in the learning curves, especially for the control group, suggests that writing symbolic expressions is a challenging skill.

Opportunity and Condition \times Opportunity interaction terms were added to the model for Unit 6 to determine if the learning curves in Figure 2 were significantly different. Results showed that for hard KCs, there was a significant and positive Condition \times Opportunity interaction ($z = 2.09$, $p = .036$) in Unit 6. This suggests that as students were given more opportunities to master hard KCs, receiving personalization incrementally improved performance above the control condition. Personalization doubled the raw gain students saw from each practice opportunity—with each opportunity, odds of a correct response increased by 1.04 for control group and 1.08 for the experimental group. Thus, results suggest that in Unit 6, students in the experimental group were able to learn expression writing with less practice. The difference between the learning curves for hard KCs in Unit 10 was not significant ($p = .79$).

Performance for struggling students. The impact of personalization was also examined based on student prior achievement.

Within the CTA environment, as previously mentioned, one of the best measures of achievement is curriculum progress. An indicator variable was added to the model to identify students who did not reach Unit 6 until halfway through the data collection period or later (February–May). Students in these classes began working in CTA in September and were expected to complete 6–7 units every 9 weeks as part of their Algebra I course. Students who reached Unit 6 in February or later (25 out of the 145 students; 13 control, 12 experimental) were far behind their peers and struggling to meet course expectations. As personalization seemed to have the largest and most consistent impact on hard KCs, this analysis was limited to the problem parts where students were writing expressions with slope and intercept terms.⁵ Fixed effects included which condition the student was in, whether the student was identified as a struggling student by the curriculum progress measure, and the interaction of these terms.

As can be seen from Table 6, low-achievement students in the personalization condition greatly outperformed low-achievement students in the control condition on hard KCs in Unit 6 (odds = $1.64 \times 2.56 = 4.20$, $p = .037$). The raw difference in performance for these struggling students was a 24.69% success rate on hard KCs for the control group versus a 57.90% success rate on hard KCs for the experimental group. Considering only students with low prior achievement, personalization had a very large, positive impact on performance for algebraic expression writing in Unit 6. However, the table also shows that there was a high variation for this effect, suggesting that personalization did not have a stable impact for all students identified in this manner. This may be a

⁴ The mean number of months it took students to reach Unit 10 from Unit 6 was 1.01 months for the control group and 1.00 months for the experimental group. This difference did not approach significance ($p = .90$).

⁵ Analyses using the full data set yielded the same results.

Table 4
Output for Hierarchical Logistic Regression Model of Performance Within Unit 6

Fixed effects	Raw coefficient	SE	Odds	z value	Significance
Intercept	−0.27	0.23	0.76	−1.21	
Condition-control	Ref.				
Condition-experimental	0.42	0.11	1.52	3.78	***
KC-hard	Ref.				
KC-medium	1.33	0.06	3.78	22.21	***
KC-easy	2.18	0.06	8.84	35.69	***
Condition-experimental:KC-medium	−0.26	0.08	0.77	−3.18	**
Condition-experimental:KC-easy	−0.07	0.09	0.94	−0.77	

Note. The second column gives raw coefficients for each predictor, which are in logit form, while the fourth column transforms the coefficients into odds, a common measure of effect size for logistic regression. Random effects are not included for brevity; however, in general, most of the variance was at the student and item levels. Hard KCs are the reference in all tables. KC = knowledge component; Ref. = reference category.

** $p < .01$. *** $p < .001$.

result of curriculum progress being only a rough proxy for achievement. Personalization still had a significant and sizeable positive impact on performance for other students (odds = 1.64, $p = .006$), so the effectiveness of the intervention was not being driven entirely by the low-achievement students.

When considering the data for low-achievement students in Unit 10, it is important to note that the overall magnitude of the differences for hard KCs in Unit 10 was smaller than in Unit 6—the performance of the control group was 39.67% correct ($0.66/[1 + 0.66]$), while the performance of the experimental group was 46.18% correct ($[(0.66 \times 1.30)/[1 + 0.66 \times 1.30]]$). However, the impact of having received the treatment might have been larger for the group of students identified as having low achievement. Unfortunately, since the classification of struggling student was based on curriculum progress, only six of the 25 originally identified struggling students made it to Unit 10 before the school year ended. The fact that 19 of the low-achievement students for whom personalization was most helpful were omitted from this analysis may explain why the magnitude of the difference was smaller in Unit 10.

Time Measures in Units 6 and 10

Learning efficiency. An analysis of the time it took students to solve problem parts in Units 6 and 10 was conducted to examine

how personalization impacts learning efficiency. The dependent quantity in the model was the number of seconds the student spent answering the problem part, measured as the time between when the problem came up on the screen and when the student finished entering an answer.

Students in the personalization condition spent significantly less time (6.93-s reduction, $p = .048$) solving problem parts involving hard KCs in Unit 6 (see Table 7). Students in the experimental group had 0.59 correct answers per minute on hard KCs, while students in the control group had 0.42 correct answers per minute on hard KCs. There were no significant time differences for easy ($p = .28$) or medium KCs ($p = .12$). A comparison of reading times within Unit 6 was also conducted by using the elapsed time between when the problem first appeared and when the student entered his or her first response as the dependent variable in the regression model. Results showed that students who received personalization spent significantly less time reading problems (7.04-s reduction, $p = .025$).

Participants' time measures in Unit 10 were examined to see if receiving personalization in Unit 6 accelerated future learning in Unit 10. Students who had received personalization in Unit 6 were still faster at writing algebraic expressions in Unit 10 (6.26-s reduction, $p = .004$; see Table 8). In Unit 10, the control group achieved 0.35 correct responses per minute on hard KCs, while the

Table 5
Output for Hierarchical Logistic Regression Model of Performance Within Unit 10—Experimental Condition Received Personalization Treatment in Unit 6

Fixed effects	Raw coefficient	SE	Odds	z value	Significance
(Intercept)	−0.42	0.18	0.66	−2.31	*
Opportunities in Unit 6	−0.022	0.004	0.98	−5.54	***
Condition-control	Ref.				
Condition-experimental	0.27	0.10	1.30	2.59	**
KC-hard	Ref.				
KC-medium	1.03	0.07	2.80	14.86	***
KC-easy	2.41	0.07	11.13	37.03	***
Condition-experimental:KC-medium	−0.15	0.10	0.86	−1.53	
Condition-experimental:KC-easy	−0.15	0.09	0.86	−1.72	

Note. Number of opportunities was median centered. KC = knowledge component; Ref. = reference category.

* $p < .05$. ** $p < .01$. *** $p < .001$.

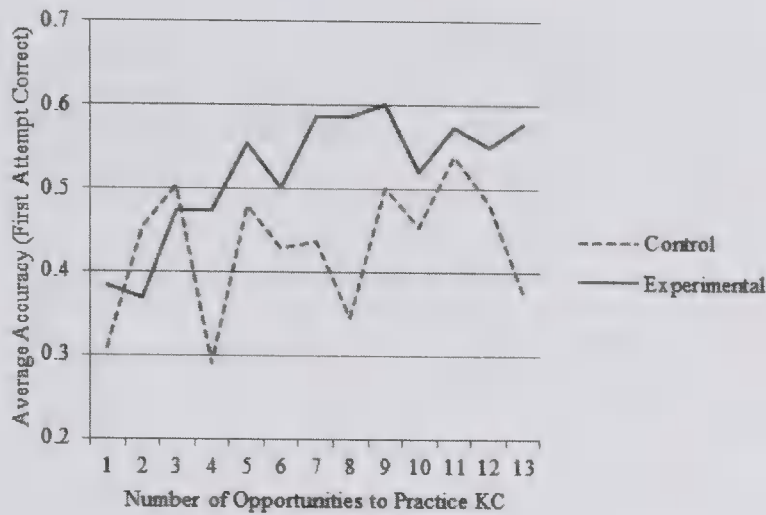


Figure 2. Learning curves for experimental and control groups on hard knowledge components (KCs) in Unit 6.

experimental group achieved 0.44 correct responses per minute on hard KCs. There were no significant time differences between experimental and control groups for easy KCs ($p = .92$) or medium KCs ($p = .23$) in Unit 10.

Gaming the system. In ITS systems, students may engage in gaming-the-system behaviors where they take advantage of hints and feedback. Baker and de Carvalho (2008) developed a gaming detector that uses measures such as the elapsed time between responses and hint requests to estimate the gaming tendency of each student. The gaming detector was run on the log data from Unit 6, and the gaming estimates for each student in the control group were compared to those in the experimental group using a Student's t test. Results showed that students in the experimental group engaged in gaming behaviors significantly less often than students the control group ($t = -2.33$, $p = .037$, Cohen's $d = 0.35$). There were no significant gaming differences in Unit 10 ($p = .737$).

Discussion

Here, we discuss the results as they correspond to each of the three research questions.

Research Questions

R.1: How does context personalization impact performance measures while the intervention is in place? In particular, how is this effect moderated by (a) the ability of the learner and (b) the difficulty of the task? Results showed that context personalization had a significant positive effect on students' performance

on easy and hard KCs in Unit 6. Personalization allowed students to more successfully write algebraic expressions from story problems when the expression included both a slope and an intercept term. This improved accuracy was reflected in students' learning curves—students who received personalization were able to master expression writing after fewer practice opportunities. This suggests that personalization may allow students to more easily compose the different terms in an algebraic expression (e.g., Hefernan & Koedinger, 1997; Koedinger & McLaughlin, 2010) and make meaning of a grounded, concrete story scenario's relation to an algebraic equation. Personalization was most beneficial for students who were struggling to progress within the CTA curriculum. Personalization may act as a support for students struggling to learn formal representational systems who are most in need of perceptual grounding. This is consistent with Mayer's (2001) *individual differences principle*, which states that design effects are stronger for low-knowledge learners because high-knowledge learners are better able to use prior knowledge to compensate for fewer supports within the environment.

We hypothesize that personalization may help to ground abstract symbols in concrete experience, allowing them to gain situation-based meaning. Writing an algebraic expression from a story is a complicated skill that can require both the construction of an accurate and meaningful situation model and the successful coordination of situation and problem models (Nathan et al., 1992). These actions may require a deep level of processing of the story situation and its quantities and relationships. The finding that personalization improves such deep-level processing measures supports Renninger et al.'s (2002) conclusion that interest-based interventions in mathematics allow learners to build important connections between the context and the mathematical content. This suggests that the personalization enhancements were not seductive details that distracted learners (Clark & Mayer, 2003) or interfered with transfer to nonpersonalized problems or abstract representations (Sloutsky, Kaminski, & Heckler, 2005). Rather, personalization may have allowed participants to better learn and understand the underlying skill of writing expressions from story scenarios, which involves successful coordination of situation-based reasoning with abstract models.

R.2: How does context personalization impact time measures while the intervention is in place? We found that context personalization allowed students to spend less time reading and solving story problems. The reduced durations seemed targeted toward one task in particular—writing symbolic equations. Personalization also decreased the incidence of gaming-the-system behaviors, or instances where students would enter answers or

Table 6
Output for Hierarchical Logistic Regression Model of Performance on Hard KCs Within Unit 6

Fixed effects	Raw coefficient	SE	Odds	z value	Significance
(Intercept)	−0.33	0.40	0.72	−0.83	
Condition-control	Ref.				
Condition-experimental	0.50	0.18	1.64	2.78	**
Regular student	Ref.				
Struggling student	−0.78	0.30	0.48	−2.64	**
Condition-experimental:struggling student	0.94	0.45	2.56	2.09	*

Note. KC = knowledge component; Ref. = reference category.

* $p < .05$. ** $p < .01$.

Table 7
Output for Hierarchical Linear Regression Model of Step Duration Within Unit 6

Fixed effects	Estimate	SE	t value	Significance
(Intercept)	57.90	7.00	8.27	***
Condition-control	Ref.			
Condition-experimental	-6.93	3.42	-2.03	*
KC-hard	Ref.			
KC-medium	-18.81	1.90	-9.91	***
KC-easy	-42.37	1.89	-22.46	***
Condition-experimental:KC-medium	3.16	2.63	1.20	
Condition-experimental:KC-easy	4.35	2.61	1.66	

Note. KC = knowledge component; Ref. = reference category.

* $p < .05$. *** $p < .001$.

request hints rapidly. We hypothesize that along with providing grounding for key ideas in algebra, personalization has an important impact on attention and engagement. Interest-based interventions may lower reaction times by allowing for automatic allocation of attention (McDaniel et al., 2000). Writing expressions involves the learner working directly with the context of the story problem and engaging in the difficult coordination between a situation model and a problem model. Personalization may focus attention when students engage in deep-level processing to accomplish this coordination. The reduction in gaming behaviors also suggests that personalization may orient learner attention toward extracting meaning, rather than rapidly requesting hints or feedback.

R.3: How does context personalization impact measures of robust learning once the intervention is removed? Four units after the intervention, students who had been in the experimental group still had significantly greater accuracy when writing algebraic expressions from normal story scenarios and wrote expressions in significantly less time compared to the control group. The results suggest that personalization promoted robust learning of the underlying concept of algebraic expression writing and was associated with accelerated future learning and transfer. The robust learning gains for Unit 10 challenge the notion that personalization is a crutch whose effects will not persist or transfer to the solving of nonpersonalized problems. Instead, the results suggest that personalization acts as a scaffold, providing grounding for students as they learn important skills relating to coordinating situation and problem models when writing algebraic expressions. Students may

be able to later flexibly apply these skills to normal story problems with more complex underlying algebraic expressions.

Significance

This study contributes to research on personalization and interest in several ways. First, we have illustrated the effectiveness of an interest-based intervention in a K-12 school during the course of regular instruction over an extended period. This allowed for a thorough and controlled examination of how performance and learning unfold in ecologically valid compulsory school settings where high stakes are attached to learning outcomes, especially in mathematics. This has been contrasted with research on pull-out studies with children (Renninger et al., 2002; Stacey & MacGregor, 1999; Walkington et al., 2012), children taking written exams (E. Bates & Wiest, 2004; Koedinger & Nathan, 2004), and studies with adult learners or laboratory settings (Durik & Harackiewicz, 2007; McDaniel et al., 2000; Reber et al., 2009).

Although previous research on the impact of personalization in mathematics has had mixed results (e.g., Cakir & Simsek, 2010; Chen & Liu, 2007), here we have shown that an interest-based intervention can be effective for mathematics learning. We hypothesized that the mixed results in the literature may reflect a need to closely match problem difficulty and student ability. We thus took a novel approach, implementing personalization in the context of an ITS that adapts instruction to current knowledge states. This study demonstrates the powerful synergistic effect of

Table 8
Output for Hierarchical Linear Regression Model of Step Duration Within Unit 10—Experimental Condition Received Personalization Treatment in Unit 6

Fixed effects	Estimate	SE	t value	Significance
(Intercept)	67.49	2.76	24.45	***
Condition-control	Ref.			
Condition-experimental	-6.26	2.16	-2.90	**
KC-hard	Ref.			
KC-medium	-34.17	1.38	-24.79	***
KC-easy	-53.38	1.28	-41.77	***
Condition-experimental:KC-medium	8.08	1.91	4.24	***
Condition-experimental:KC-easy	6.35	1.77	3.59	***

Note. KC = knowledge component; Ref. = reference category.

** $p < .01$. *** $p < .001$.

using technology systems to adapt instruction to both knowledge states and personal interests.

Furthermore, as previous research on interest has largely focused on domains like reading or arithmetic, we contribute to an understanding of the ways in which interest can mediate learning in an advanced domain that emphasizes abstract representations, relational thinking, and generalization. Research on the development of algebraic reasoning has accentuated the importance of situating algebra learning in rich, engaging mathematical contexts (Bardini et al., 2004; Carraher et al., 2006); however, students often find secondary courses meaningless with respect to their everyday experiences (Mitchell, 1993). Here, we have shown that even minor modifications to make problems more interest based can have a positive impact on learning. We hypothesized that personalization may allow students to learn *how* to construct meaningful situation models of stories and coordinate them with problem models by providing perceptual grounding that makes abstractions more meaningful. Coordinating situation-based reasoning with formal mathematical modeling is a critical skill not only for algebraic expression writing but for many important applications of mathematics (e.g., Common Core State Standards Initiative, 2010).

Future Directions

Discussion of mediators in the relationship between personalization and improved learning in the present study is speculative, as these constructs were not directly measured. In future work, we plan to explore the mediators through which personalization is associated with learning more directly. We will track at a fine-grained level how personalization interacts with measures of motivational and metacognitive variables (e.g., situational interest, self-efficacy, etc.) using questionnaires (see Bernacki, Nokes-Malach, & Aleven, in press). Recent advances in technology-based affective state detection also offer an exciting direction for measuring affect without relying solely on self-report data. We plan to incorporate into our intervention affective state detectors (Baker et al., 2012) that conduct computational analyses of response behaviors to detect states like engagement, boredom, or frustration. In this way, we plan to further delineate the path between a personalization intervention and improved robust learning.

Conclusion

The future of adaptive learning technologies in classrooms is both promising and generative—the National Academy of Engineering recently named the development of personalized systems for learning as one of the grand challenges for engineering in the 21st century (Ellis, 2009). Here, we have shown the potential of a simple, interest-based adaptation in promoting student learning. However, as technology advances, more powerful methods will emerge to customize learning to the events, objects, and activities that are personally relevant, evocative, and motivating to students in K-12 schools today. Personalization systems should be designed to leverage students' interests in authentic and meaningful ways, creating systems that allow learners to collaborate around complex open-ended tasks that are situated within and adapted to their experiences. Such tasks can provide an authentic venue for abstract systems of representation to arise as powerful tools for modeling

the world and making sense of personally relevant phenomena. Here, we have shown only a hint of what the potential of such a system could be. However, with this work, we contribute to the foundation for what we believe the designers of adaptive learning technologies should look toward. Students today are accustomed to customization, interaction, and control when seeking knowledge—modern learning environments and those who design them must themselves adapt to the rapid technological changes taking place in our world.

References

- Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94*, 545–561. doi:10.1037/0022-0663.94.3.545
- Ainley, M., Hillman, K., & Hidi, S. (2002). Gender and interest processes in response to literary texts: Situational and individual interest. *Learning & Instruction, 12*, 411–428.
- Anand, P., & Ross, S. (1987). Using computer-assisted instruction to personalize arithmetic materials for elementary school children. *Journal of Educational Psychology, 79*, 72–78. doi:10.1037/0022-0663.79.1.72
- Baker, R., Corbett, A., Koedinger, K., & Wagner, A. (2004). Off-task behavior in the cognitive tutor classroom: When students “game the system”. In E. Dykstra-Erickson & M. Tscheligi (Eds.), *Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems* (pp. 383–390). New York, NY: ACM Press.
- Baker, R., & de Carvalho, A. (2008). Labeling student behavior faster and more precisely with text replays. In R. S. J. de Baker, T. Barnes, & J. E. Beck (Eds.), *Educational data mining: 1st International Conference on Educational Data Mining, Proceedings* (pp. 38–47). Montreal, Quebec, Canada: International Educational Data Mining Society.
- Baker, R., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., . . . Rossi, L. (2012). Sensor-free automated detection of affect in a Cognitive Tutor for Algebra. In K. Yacef, O. Zaiane, H. HersHKovitz, M. Yudelson, & J. Stamper (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 126–133). Chania, Greece: International Educational Data Mining Society.
- Bardini, C., Pierce, R., & Stacey, K. (2004). Teaching linear functions in context with graphics calculators: Students' responses and the impact of the approach on their use of algebraic symbols. *International Journal of Science and Mathematics Education, 2*, 353–376. doi:10.1007/s10763-004-8075-3
- Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes—R package version 0.999375–35 [Computer software]. Retrieved from <http://CRAN.R-project.org/package=lme4>
- Bates, E., & Wiest, L. (2004). The impact of personalization of mathematical word problems on student performance. *Mathematics Educator, 14*(2), 17–26.
- Bernacki, M. L., Nokes-Malach, T. J., & Aleven, V. (in press). Fine-grained assessment of motivation over long periods of learning with an intelligent tutoring system: Methodology, advantages, and preliminary results. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies*. Berlin, Germany: Springer.
- Breidenbach, D., Dubinsky, E., Hawks, J., & Nichols, D. (1992). Development of the process conception of function. *Educational Studies in Mathematics, 23*, 247–285. doi:10.1007/BF02309532
- Cakir, O., & Simsek, N. (2010). A comparative analysis of computer and paper-based personalization on student achievement. *Computers & Education, 55*, 1524–1531. doi:10.1016/j.compedu.2010.06.018
- Carnegie Learning. (2011). Carnegie Learning math series: Carnegie Learning MATHia Software [Computer software]. Retrieved from <http://mathseries.carnegielearning.com/product-info/software>

- Carraher, D., Schliemann, A., Brizuela, B., & Earnest, D. (2006). Arithmetic and algebra in early mathematics education. *Journal for Research in Mathematics Education*, 37, 87–115.
- Chazan, D. (1999). On teachers' mathematical knowledge and student exploration: A personal story about teaching a technologically supported approach to school algebra. *International Journal of Computers for Mathematical Learning*, 4, 121–149. doi:10.1023/A:1009875213030
- Chen, C., & Liu, P. (2007). Personalized computer-assisted mathematics problem-solving program and its impact on Taiwanese students. *Journal of Computers in Mathematics and Science Teaching*, 26, 105–121.
- Clark, R. C., & Mayer, R. E. (2003). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. San Francisco, CA: Jossey-Bass/Pfeiffer.
- Clement, J. (1982). Algebra word problem solutions: Thought processes underlying a common misconception. *Journal for Research in Mathematics Education*, 13, 16–30. doi:10.2307/748434
- Collins, A., & Halverson, R. (2009). *Rethinking education in the age of technology: The digital revolution and schooling in America*. New York, NY: Teachers College Press.
- Common Core State Standards Initiative. (2010). *Common Core State Standards (Mathematics Standards)*. Retrieved from <http://www.corestandards.org/the-standards/mathematics>
- Cordova, D., & Lepper, M. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88, 715–730. doi:10.1037/0022-0663.88.4.715
- Davis-Dorsey, J., Ross, S., & Morrison, G. (1991). The role of rewording and context personalization in the solving of mathematical word problems. *Journal of Educational Psychology*, 83, 61–68. doi:10.1037/0022-0663.83.1.61
- Durik, A., & Harackiewicz, J. (2007). Different strokes for different folks: How individual interest moderates effects of situational factors on task interest. *Journal of Educational Psychology*, 99, 597–610. doi:10.1037/0022-0663.99.3.597
- Ellis, G. (2009). Grand challenges for engineering. *IEEE Engineering Management Review*, 37(1), 3. doi:10.1109/EMR.2009.4804341
- Filloy, E., & Rojano, T. (1989). Solving equations: The transition from arithmetic to algebra. *For the Learning of Mathematics*, 9(2), 19–25.
- Flowerday, T., Schraw, G., & Stevens, J. (2004). The role of reader choice and interest in reader engagement. *Journal of Experimental Education*, 72, 93–114. doi:10.3200/JEXE.72.2.93-114
- Fredricks, J. A., & Eccles, J. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology*, 38, 519–533.
- Frenzel, A., Goetz, T., Pekrun, R., & Watt, H. (2010). Development of mathematics interest in adolescence: Influences of gender, family, and school context. *Journal of Research on Adolescence*, 20, 507–537. doi:10.1111/j.1532-7795.2010.00645.x
- Goldstone, R., & Son, J. (2005). The transfer of scientific principles using concrete and idealized simulations. *Journal of the Learning Sciences*, 14, 69–110. doi:10.1207/s15327809jls1401_4
- Harackiewicz, J., Durik, A., Barron, K., Linnenbrink, E., & Tauer, J. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology*, 100, 105–122. doi:10.1037/0022-0663.100.1.105
- Harackiewicz, J., Rozek, C., Hulleman, C., & Hyde, J. (2012). Helping parents to motivate adolescents in math and science: An experimental test of a utility value intervention. *Psychological Science*, 23, 899–906.
- Heffernan, N. T., & Koedinger, K. R. (1997). The composition effect in symbolizing: The role of symbol production vs. text comprehension. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society* (pp. 307–312). Mahwah, NJ: Erlbaum.
- Hegarty, M., Mayer, R., & Monk, C. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, 87, 18–32. doi:10.1037/0022-0663.87.1.18
- Heilman, M., Collins, K., Eskenazi, M., Juffs, A., & Wilson, L. (2010). Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20, 73–98.
- Herscovics, N., & Linchevski, L. (1994). A cognitive gap between arithmetic and algebra. *Educational Studies in Mathematics*, 27, 59–78. doi:10.1007/BF01284528
- Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research*, 60, 549–571.
- Hidi, S. (2001). Interest, reading and learning: Theoretical and practical considerations. *Educational Psychology Review*, 13, 191–209. doi:10.1023/A:1016667621114
- Hidi, S., & Ainley, M. (2008). Interest and self-regulation. The relationships between two variables that influence learning. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and application* (pp. 77–109). New York, NY: Erlbaum.
- Hidi, S., & Harackiewicz, J. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, 70, 151–179.
- Hidi, S., & Renninger, K. (2006). The four-phase model of interest development. *Educational Psychologist*, 41, 111–127. doi:10.1207/s15326985ep4102_4
- Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology*, 102, 880–895. doi:10.1037/a0019506
- Hulleman, C. S., & Harackiewicz, J. M. (2009, December 4). Promoting interest and performance in high school science classes. *Science*, 326, 1410–1412. doi:10.1126/science.1177067
- Humberstone, J., & Reeve, R. (2008). Profiles of algebraic competence. *Learning and Instruction*, 18, 354–367. doi:10.1016/j.learninstruc.2007.07.002
- Kaput, J. J. (2000). *Teaching and learning a new algebra with understanding*. Dartmouth, MA: National Center for Improving Student Learning and Achievement in Mathematics and Science.
- Kintsch, W. (1986). Learning from text. *Cognition and Instruction*, 3, 87–108. doi:10.1207/s1532690xc0302_1
- Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review*, 19, 239–264. doi:10.1007/s10648-007-9049-0
- Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–77). St. Louis, MO: Cambridge University Press.
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction (KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36, 1–42. doi:10.1111/j.1551-6709.2012.01245.x
- Koedinger, K., & McLaughlin, E. (2010). Seeing language learning inside the math: Cognitive analysis yields transfer. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 471–476). Portland, OR: Cognitive Science Society.
- Koedinger, K., & Nathan, M. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences*, 13, 129–164. doi:10.1207/s15327809jls1302_1
- Ku, H., & Sullivan, H. (2000). Personalization of mathematics word problems in Taiwan. *Educational Technology Research and Development*, 48, 49–60. doi:10.1007/BF02319857
- Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven, CT: Yale University Press.

- Linnenbrink-Garcia, L., Durik, A., Conley, A., Barron, K., Tauer, J., Karabenick, S., & Harackiewicz, J. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement*, 70, 647–671. doi:10.1177/0013164409355699
- López, C. L., & Sullivan, H. J. (1992). Effects of personalization of instructional context on the achievement and attitudes of Hispanic students. *Educational Technology Research and Development*, 40, 5–14. doi:10.1007/BF02296895
- Mathan, S., & Koedinger, K. (2005). Fostering the intelligent novice: Learning with errors from metacognitive tutoring. *Educational Psychologist*, 40, 257–265. doi:10.1207/s15326985ep4004_7
- Mayer, R. E. (2001). *Multimedia learning*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9781139164603
- McCoy, L. P. (2005). Effect of demographic and personal variables on achievement in eighth-grade algebra. *Journal of Educational Research*, 98(3), 131–135.
- McDaniel, M., Waddill, P., Finstad, K., & Bourg, T. (2000). The effects of text-based interest on attention and recall. *Journal of Educational Psychology*, 92, 492–502. doi:10.1037/0022-0663.92.3.492
- Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology*, 85, 424–436. doi:10.1037/0022-0663.85.3.424
- Moses, R., & Cobb, C. (2001). *Radical equations: Math literacy and civil rights*. Boston, MA: Beacon Press.
- Nathan, M., Kintsch, W., & Young, E. (1992). A theory of algebra-word problem comprehension and its implications for the design of learning environments. *Cognition and Instruction*, 9, 329–389. doi:10.1207/s1532690xc0904_2
- Papert, S. (1980). *Mindstorms: Children, computers and powerful ideas*. New York, NY: Basic Books.
- Papert, S. (1993). *The children's machine*. New York, NY: Harper Collins.
- Reber, R., Hetland, H., Chen, W., Norman, E., & Kobbeltvedt, T. (2009). Effects of example choice on interest, control, and learning. *Journal of the Learning Sciences*, 18, 509–548. doi:10.1080/10508400903191896
- Renninger, K., Ewen, L., & Lasher, A. (2002). Individual interest as context in expository text and mathematical word problems. *Learning and Instruction*, 12, 467–490. doi:10.1016/S0959-4752(01)00012-3
- Renninger, K., & Hidi, S. (2011). Revisiting the conceptualization, measurement, and generation of interest. *Educational Psychologist*, 46, 168–184. doi:10.1080/00461520.2011.587723
- Renninger, K., & Su, S. (2012). Interest and its development. In R. M. Ryan (Ed.), *The Oxford handbook of human motivation* (pp. 167–189). New York, NY: Oxford University Press.
- Renninger, K., & Wozinak, R. (1985). Effect of interest on attentional shift, recognition, and recall in young children. *Developmental Psychology*, 21, 624–632. doi:10.1037/0012-1649.21.4.624
- Sansone, C., Fraughton, T., Zachary, J., Butner, J., & Heiner, C. (2011). Self-regulation of motivation when learning online: The importance of who, why and how. *Educational Technology Research and Development*, 59, 199–212. doi:10.1007/s11423-011-9193-6
- Schiefele, U. (1990). The influence of topic interest, prior knowledge, and cognitive capabilities on text comprehension. In J. M. Pieters, K. Breuer, & P. R. J. Simons (Eds.), *Learning environments* (pp. 323–338). Berlin, Germany: Springer. doi:10.1007/978-3-642-84256-6_25
- Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist*, 26, 299–323.
- Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading*, 3, 257–279. doi:10.1207/s1532799xssr0303_4
- Schmidt, R., & Bjork, R. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3, 207–217. doi:10.1111/j.1467-9280.1992.tb00029.x
- Sloutsky, V. M., Kaminski, J. A., & Heckler, A. F. (2005). The advantage of simple symbols for learning and transfer. *Psychonomic Bulletin & Review*, 12, 508–513. doi:10.3758/BF03193796
- Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.
- Stacey, K., & MacGregor, M. (1999). Learning the algebraic method of solving problems. *Journal of Mathematical Behavior*, 18, 149–167. doi:10.1016/S0732-3123(99)00026-7
- Swafford, J., & Langrall, C. (2000). Grade 6 students' preinstructional use of equations to describe and represent problem situations. *Journal for Research in Mathematics Education*, 31, 89–112. doi:10.2307/749821
- Walkington, C., Petrosino, A., & Sherman, M. (in press). Supporting algebraic reasoning through personalized story scenarios: How situational understanding mediates performance and strategies. *Mathematical Thinking and Learning*.
- Walkington, C., Sherman, M., & Petrosino, A. (2012). "Playing the game" of story problems: Coordinating situation-based reasoning with algebraic representation. *Journal of Mathematical Behavior*, 31, 174–195. doi:10.1016/j.jmathb.2011.12.009
- Wilensky, U. (1991). Abstract meditations on the concrete and concrete implications for mathematics education. In I. Harel & S. Papert (Eds.), *Constructionism* (pp. 193–203). Norwood, NJ: Ablex Publishing.

Received December 15, 2011

Revision received September 4, 2012

Accepted December 18, 2012 ■

Generalizing Automated Detection of the Robustness of Student Learning in an Intelligent Tutor for Genetics

Ryan S. J. d. Baker
Teachers College, Columbia University

Albert T. Corbett
Carnegie Mellon University

Sujith M. Gowda
Worcester Polytechnic Institute

Recently, there has been growing emphasis on supporting robust learning within intelligent tutoring systems, assessed by measures such as transfer to related skills, preparation for future learning, and longer term retention. It has been shown that different pedagogical strategies promote robust learning to different degrees. However, the student modeling methods embedded within intelligent tutoring systems remain focused on assessing basic skill learning rather than robust learning. Recent work has proposed models, developed using educational data mining, that infer whether students are acquiring learning that transfers to related skills, and prepares the student for future learning (PFL). In this earlier work, evidence was presented that these models achieve superior prediction of robust learning to what can be achieved by traditional methods for student modeling. However, using these models to drive intervention by educational software depends on evidence that these models remain effective within new populations. To this end, we analyze the degree to which these detectors remain accurate for an entirely new population of high school students. We find limited evidence of degradation for transfer. More degradation is seen for PFL. This degradation appears to occur in part because it is generally more difficult to infer this construct within the new population.

Keywords: robust learning, preparation for future learning, transfer, student modeling, intelligent tutoring system

Increasingly, it is thought desirable that students acquire what is termed *robust* knowledge (Koedinger, Corbett, & Perfetti, 2012): knowledge grounded in conceptual domain knowledge (Craig, VanLehn, & Chi, 2008), which transfers more readily to related problem situations (Fong & Nisbett, 1991; Singley & Anderson, 1989), is retained by students over time (Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Schmidt & Bjork, 1992), and prepares

students for more efficient or more effective future learning (Bransford & Schwartz, 1999; Schwartz & Martin, 2004). One of the well-documented risks in problem solving across STEM (science, technology, engineering, and mathematics) domains is that students can develop superficial knowledge that fails these tests of robust learning. In particular, when students are not well prepared for problem solving, they can develop problem-solving knowledge that focuses on surface elements in problem situations, formal representations, and features of the learning environment itself (Chi, Feltovich, & Glaser, 1981; Rittle-Johnson & Siegler, 1998).

In line with this shift in perspective, over the past 15 years there has been a growing effort by intelligent tutoring system (ITS) developers and developers of other intelligent learning environments (ILEs) to develop interventions explicitly designed to increase the robustness of student learning. One general theme has been to improve the effectiveness of tutor feedback in supporting deep understanding, for example, through natural language tutorial dialogues (Graesser et al., 2004; Katz, Connelly, & Wilson, 2007), through enhanced student interactivity with graphical feedback (Butcher, 2010; Corbett & Trask, 2000), or through focusing feedback on domain-independent strategies (Chi & VanLehn, 2007). A second major approach has focused on incorporating student explanations into ITSs, asking students to explain their actions in problem solving (Aleven & Koedinger, 2002), or to explain worked examples of problem solutions (Corbett et al., 2011; Hausmann & VanLehn, 2007; McLaren, Lim, & Koedinger, 2008; Schwonke et al., 2009), toward supporting students in mon-

This article was published Online First September 9, 2013.

Ryan S. J. d. Baker, Department of Human Development, Teachers College, Columbia University; Albert T. Corbett, Department of Psychology, Carnegie Mellon University; Sujith M. Gowda, Department of Social Science and Policy Studies, Worcester Polytechnic Institute.

This research was supported via grant "Empirical Research: Emerging Research: Robust and Efficient Learning: Modeling and Remediating Students' Domain Knowledge"; National Science Foundation (NSF) Award DRL-0910188, via grant "Promoting Robust Understanding of Genetics With a Cognitive Tutor That Integrates Conceptual Learning With Problem Solving"; Institute of Education Sciences Award R305A090549; and by the Pittsburgh Science of Learning Center, NSF Award SBE-0836012. Portions of the work published here were previously presented in conference papers (Baker, Gowda, & Corbett, 2011a, 2011b). We thank Robert Siegler, Ken Koedinger, Kurt VanLehn, Ben MacLaren, Lisa Rossi, and Belinda Yew for helpful comments and suggestions.

Correspondence concerning this article should be addressed to Ryan S. J. d. Baker, Department of Human Development, Teachers College, Columbia University, 525 West 120th Street, New York, NY 10522. E-mail: baker2@exchange.tc.columbia.edu

itoring their understanding. Other efforts have focused on training metacognitive skills, such as the skill of using a tutoring system's corrective and explanatory feedback effectively (Aleven, McLaren, Roll, & Koedinger, 2006; Roll, Aleven, McLaren, & Koedinger, 2007), and providing meta-cognitive feedback on students' skill at self-regulated learning (Chin et al., 2010; Tan & Biswas, 2006).

The advent of interventions that can support the development of robust learning raises the question of whether another major benefit of intelligent tutors and artificial intelligence in education (AIED) technologies can be leveraged: individualization. Individualization is a major goal of ITS and AIED systems (cf. McCalla, 1992; VanLehn, 2006), driven by models of students' latent knowledge (cf. Corbett & Anderson, 1995; Martin & VanLehn, 1995; Shute, 1995). Individualization based on student knowledge has had substantial benefits for learners. For instance, Corbett (2001) demonstrated that Bayesian student modeling can be used to more efficiently distribute problem-solving practice in an ITS, leading to a large gain in mean posttest accuracy with only a small additional cost in total time on task, compared with a fixed curriculum. Bayesian student modeling has also been successfully used to monitor student explanations of worked examples in ITSs (Conati, Gertner, & VanLehn, 2002; Salden, Koedinger, Renkl, Aleven, & McLaren, 2010).

Efforts to individualize learning environments rely on accurate student modeling. The efforts listed above have leveraged models of student knowledge that can successfully infer the probability that a student knows a specific skill from the student's history of correct responses and noncorrect responses (e.g., errors and hint requests) for that skill up until that time (cf. Corbett & Anderson, 1995; Martin & VanLehn, 1995; Pavlik, Cen, & Koedinger, 2009; Shute, 1995). In recent years, the debate about how to best model student knowledge has continued, with an increasing number of explicit comparisons of models' ability to predict future performance within the tutoring software studied (cf. Gong, Beck, & Heffernan, 2010; Pardos, Gowda, Baker, & Heffernan, 2011; Pavlik et al., 2009; Wang & Heffernan, 2011).

Although these student modeling approaches have been successful at predicting immediate problem-solving performance and improving performance on those tests, less attention has been paid to modeling the robustness of student learning. Several studies have shown that Bayesian student modeling can accurately predict immediate posttest performance on the same problem-solving skills studied with a tutor (e.g., Baker et al., 2010; Corbett & Anderson, 1995; Corbett, McLaren, Kauffman, Wagner, & Jones, 2010; Pardos et al., 2011; Shute, 1995), a very limited form of transfer. But student models in ITSs have typically not attempted to go beyond this point in modeling whether learning is robust. Relatedly, some results suggest that Bayesian student modeling can be insensitive to differences in students' depth of understanding. For example, Corbett and Anderson (1995) reported that whereas Bayesian student modeling achieved high correlation to student posttest performance in the APT Lisp Tutor, it overestimated average student posttest performance by 5%–10%. Tellingly, Corbett and Bhatnagar (1997) found that the extent to which the student model overestimates student test performance is inversely correlated with the each student's initial declarative knowledge. In another APT Lisp Tutor study (Corbett & Trask, 2000), two groups of students worked to cognitive mastery levels with

conventional and enhanced feedback related to a difficult topic. Although students in the two groups worked to the same nominal cognitive mastery criterion, students in the enhanced feedback condition scored reliably better on the posttest, again suggesting that this type of student modeling may be partially insensitive to differences in deep understanding.

Some steps in the direction of modeling the robustness of learning in ITSs have been taken. For example, Jastrzembski, Gluck, and Gunzelmann (2006) predict not just posttest performance, but also how long knowledge will be retained after learning, within an ITS teaching flight skills. Another step in this direction is to assess the transfer of skill within the learning system. Much of this work has taken the form of modeling interconnections between skills during learning (cf. Martin & VanLehn, 1995) or online testing (Desmarais, Meshkinfam, & Gagnon, 2006), or in using interconnections between skills to revise skill models (Pavlik, Cen, Wu, & Koedinger, 2008). Additional, computational modeling has analyzed the mechanisms leading to accelerated future learning within a learning system (Li, Cohen, & Koedinger, 2010).

Building on this work, recent work has used data mining to develop models that can automatically detect whether student knowledge will transfer to related skills outside of the tutoring system, and whether students are prepared for future learning outside of the tutoring system. The difference between transfer and PFL is whether students have the ability to directly apply their existing knowledge in novel situations or in new fashions (transfer), versus whether students can acquire new knowledge more quickly or effectively from future instruction, using their existing knowledge (preparation for future learning [PFL]). If models are developed that accomplish these goals—predicting from in-tutor behavior whether a student will be able to successfully transfer her or his knowledge out of the tutor to different skills and situations, and whether a student will be prepared for future learning outside of the tutor—then these models could be used to identify students who may be developing superficial knowledge in problem solving and in selecting interventions designed to improve the robustness of student learning. Students who are already on the road to robust learning could continue with existing activities, whereas students unlikely to achieve robust learning could receive interventions.

In this earlier work, robust learning detectors (for both transfer and PFL) were developed for a population of undergraduate students using a Cognitive Tutor in the domain of Genetics problem solving (Corbett et al., 2010). These detectors were generated by engineering complex features related to students' motivation and metacognition and creating a model to predict transfer/PFL from these features. They were assessed using cross-validation at the student level (e.g., the detectors were repeatedly developed using data from one group of students and tested on other students). The detectors were found to be better than traditional student modeling methods for predicting both transfer and PFL. In this article, we study how well these detectors of transfer and PFL generalize at the population level, studying the degree to which they transfer to a new group of students, specifically, a younger group of high school students using the same tutor software.

In addition to examining the models' degree of generalization, we also analyze the specific student behaviors that are associated with robust learning in each population, toward increasing under-

standing of the conditions under which robust learning occurs in interactive learning systems of this type.

Learning System

Cognitive Tutors are a type of interactive learning environment in which cognitive modeling and artificial intelligence are used to model student learning, in turn using the model of student learning to adapt to individual differences in student knowledge and learning (Koedinger & Corbett, 2006). Cognitive Tutor curricula combine conceptual instruction delivered by a teacher with computer-based learning where each student works one-on-one with a Cognitive Tutoring system that chooses exercises and feedback on the basis of a running model of which skills the student possesses (Corbett & Anderson, 1995).

Within a Cognitive Tutor, as the student works through a set of problems, Bayesian knowledge tracing (Corbett & Anderson, 1995) is used to determine how well the student is learning component skills, calculating the probability that the student knows each skill based on that student's history of responses within the tutor. Using these estimates of student knowledge, the tutoring system gives each student problems that are relevant to the skills that he or she needs to learn, continuing to provide problems until the student reaches mastery (e.g., 95% probability of knowing each skill) on all skills relevant to a given curricular area.

Within this article, we study robust learning in the context of the Genetics Cognitive Tutor (Corbett et al., 2010). This tutor consists of 19 modules that support problem solving across a wide range of topics in genetics (Mendelian transmission, pedigree analysis, gene mapping, gene regulation, and population genetics). Various subsets of the 19 modules have been piloted at 15 universities in North America. This study focuses on a tutor module that uses a

gene mapping technique called a *three-factor cross* (3FC). The tutor interface for this reasoning task is displayed in Figure 1. The 3FC technique is used to determine both the order of three genes (F, G, and H in this example), which lie on one chromosome, and to find the relative distances between the pairs of genes. In this technique, two organisms are crossed (two fruit flies in the example), and the resulting distribution of offspring phenotypes is analyzed to infer the arrangement of the three genes on the chromosome. In Figure 1, the student has almost finished the problem. The student has summed the number of offspring in each of four phenotype groups that appear in the offspring table, and has categorized each group (as "parental single crossover" during meiosis, or "double crossover" during meiosis). The student has compared the phenotype patterns in the offspring groups to identify the middle of the three genes and entered a gene sequence below the table. Finally, in the lower right of Figure 1, the student has calculated the crossover frequency between two of the genes, G and H, and the distance between the two genes. The student will perform the last two steps for the other two gene pairs.

Robust Learning Measures

The robustness of student learning was measured through two tests: a transfer test and a PFL test. A standard pretest and posttest, measuring the exact skills studied in the tutor, were also given.

The transfer test consisted of two problems. The first problem was a 3FC task in which double crossovers were so improbable that the double-crossover offspring group was missing. This is a "gap filling" transfer task (cf. VanLehn, Jones, & Chi, 1992). The problem is solvable and most of the students' problem-solving knowledge directly applies; the task examines whether students can draw on their understanding of that problem-solving knowl-

Student Teacher

7. In a student lab, a test cross was performed between a fruit fly that was heterozygous for three genes and one that was homozygous recessive. The offspring were scored for the three phenotypes. The student's data is shown below. Determine the gene order and the map distances for the three genes.

0. Frequency of Offspring Types

Type	Number	Group
G H f	3	I
g h F	6	I
g H f	52	II
G h F	59	II
G H F	32	III
g h f	39	III
g H F	388	IV
G h f	421	IV

1. Classify Offspring Groups

# in Group	Offspring Type of Group
9	Parental single crossover
111	Parental single crossover
71	Double crossover
809	Parental single crossover

Total 1000

2. Order Genes on the Chromosome

Gene 1	Gene 2	Gene 3
G	H	f

3. Compute Distance between each Gene Pair

Gene Pair	Frequency of Recombination	Map Units
G - H	10.0%	10.0
G - f	10.0%	10.0
H - f	10.0%	10.0

Help Done

Figure 1. The three-factor cross lesson of the Genetics Cognitive Tutor.

edge to fill in the “gap” that results from the missing offspring group. The second problem examines whether students can extend their understanding of crossovers and crossover notation from three genes to four genes. In this problem, students were given a parental genotype with four genes and asked to identify how many crossovers had occurred in various offspring groups (based on phenotype structure rather than relative frequency) and to identify all the offspring groups in which a specific crossover had occurred. Students completed this transfer test following the problem-solving posttest at the end of Session 2.

It is worth noting that the form of transfer represented by these problems can be seen as different from simply transferring knowledge to an isomorphic problem (cf. Gick & Holyoak, 1987). However, transfer problems of the more complex nature seen here, requiring some reasoning beyond simply transfer of skill, are frequently also seen in research on robust learning in interactive learning software (cf. Aleven & Koedinger, 2002; Atkinson, 2002; Hausmann & VanLehn, 2007; Mathan & Koedinger, 2005), and may represent a deeper test of the robustness of knowledge than an isomorphic problem. Interestingly, this more complex type of transfer problem is sometimes termed *far transfer*, but it is not yet clear whether it is more difficult for students to modify their knowledge to accomplish a related task (the type of transfer seen here) or whether it is more difficult for them to realize that their existing knowledge applies in a different context (the type of transfer studied in Gick & Holyoak, 1987).

In the PFL test, students were asked to solve parts of a four-factor cross problem. The reasoning is related to solving a 3FC problem, but sufficiently more complicated that a student could not be expected to invent a solution method by direct transfer, and certainly not in a short period of time. Consequently, this PFL test presented a 2.5-page description of the reasoning in a four-factor cross experiment, then asked students to solve some elements of a four-factor cross problem: identifying the middle genes, identifying all the offspring groups with a crossover between two specific genes, finding the map distance between those two genes.

Previous Models

In Baker, Gowda, and Corbett (2011a, 2011b), we presented models that can predict student transfer and PFL. These models were developed using data from 72 college students enrolled in biology courses at Carnegie Mellon University, who used the Genetics Cognitive Tutor for 2 hr apiece. The students used the Cognitive Tutor software for 2 hr, completing a total of 22,885 problem-solving attempts across a total of 10,966 problem steps in the tutor.

Feature Engineering

The first step of our process of developing models of robust learning was to engineer a set of features on the basis of a combination of theory and prior work detecting related behaviors. We tested a set of 18 features, represented as a set of nine core features and nine related features. Features 1–5 and their related features focus on student interactions with the tutor’s hints and feedback. Features 6–8 and their related features focus on the student’s problem-solving actions. The ninth feature involves the dynamics of the student’s learning, moment by moment.

1. Help avoidance (Aleven et al., 2006), not requesting help on poorly known skills (on the student’s first attempt at a specific problem step), and a related feature, Feature 1’, not requesting help on well-known skills.

2. Long pauses after receiving bug messages (error messages given when the student’s behavior indicates a known misconception), which may indicate self-explanation (cf. Chi, Bassok, Lewis, Reimann, & Glaser, 1989) of the bug message, and its inverse, Feature 2’, short pauses after receiving bug messages (indicating a failure to self-explain).

3. Long pauses after reading on-demand help messages (potentially indicating deeper knowledge or self-explanation), and an inverse feature, Feature 3’, short pauses after reading the on-demand help message.

4. Long pauses after reading an on-demand help message and getting the current action right (cf. Shih, Koedinger, & Scheines, 2008), and an inverse feature, Feature 4’, short pauses after reading an on-demand hint message and getting the current action right. Features 4 and 4’ are subsets of Features 3 and 3’.

5. Long pauses on skills that the student probably knows (may indicate continuing to self-explain even after proceduralization), and an inverse feature, Feature 5’, short pauses on skills assessed as known.

6. Off-task behavior (Baker, 2007), where the student is engaged in behavior that does not involve the system or a learning task, and a related feature, Feature 6’, long pauses that are not off-task (may indicate self-explanation, or asking teacher for help; cf. Schofield, 1995). Off-task behavior is assessed using an automated detector (Baker, 2007).

7. Gaming the system (Baker, Corbett, Roll, & Koedinger, 2008), attempting to succeed at problem steps without learning the material (by clicking through help messages quickly until receiving the answer, or systematic guessing), and a related feature, Feature 7’, fast actions that do not involve gaming (which may indicate a very well-known skill). These features are computed using an automated detector of gaming the system (Baker, Corbett, et al., 2008).

8. The student’s average probability of contextual slip/carelessness on errors, making an error when the student is assessed to know the relevant skill (known to predict posttest problem-solving performance; Baker et al., 2010). This feature is computed using an automated detector (Baker et al., 2010). Also, a related feature, Feature 8’, the certainty of contextual slip, the average contextual slip computed only for values of contextual slip over 0.5; this represents how certain the model is when it indicates that a student has slipped.

9. The student’s average learning-per-learning opportunity using the moment-by-moment learning model, which estimates the probability that the student learned a relevant skill at each step in problem solving. Also, a related feature, Feature 9’, the degree to which there are *spikes in learning*, defined as the ratio between the maximum moment-by-moment learning and the average moment-by-moment learning.

Many of these features involve a continuous variable, such as the time taken between actions or the probability of knowing a skill. In general, our detectors do not hinge on a student’s average value for the feature (e.g., average time between actions), but instead hinge on the proportion of actions that meet a constraint (e.g., the proportion of actions with a short pause, or the proportion

of actions with a long pause). For each such feature, we empirically determined a cutoff value that indicates whether the student behavior occurred or not (e.g., a long pause or low probability), rather than averaging the actual values (times or probabilities), in order to avoid having a small proportion of extreme behaviors of interest be overwhelmed by noise in the rest of the student's data.

Once feature engineering had been completed, a three-step process was conducted to develop a model of transfers and PFL: selecting features, optimizing feature cutoffs, and combining the features into a unified prediction model. In order to select a set of features, we fit a one-parameter linear regression model predicting transfer from each feature (or related feature), using correlation as the measure of each feature's goodness. In order to increase the probability of a generalizable model, we assessed each model's correlation using student-level leave-out-one-cross-validation (LOOCV). In this approach, a model is repeatedly fit for every student except one, and then goodness of fit is tested on the left-out student. Every student is excluded from the training set and used as the test set exactly once. In this situation, each model fit can have either a positive or negative coefficient; therefore, the sign of a cross-validated correlation does not imply the direction of a relationship, but instead implies its consistency. A positive cross-validated correlation implies that the models generalize across the data, whereas a negative cross-validated correlation implies that the models fail to generalize across the data (and the relationship actually flips direction for a substantial number of students). Using cross-validation in this fashion is considered a valid alternative to statistical significance testing (cf. Raftery, 1995), which explicitly examines the goodness of the models on new data, rather than investigating how well the model fits the data it is trained on (Efron & Gong, 1983).

Transfer Detector

Only features with positive cross-validated correlation to the transfer or PFL test were considered for inclusion in the full model.

For the transfer detector, nine features met this criterion: Feature 1 (help avoidance), with a cutoff of 70% probability for "poorly known"; Feature 2 (long pauses after a bug message), with a cutoff of 7 s for "long"; Feature 2' (short pauses after a bug message), with a cutoff of 1.5 s for "long"; Feature 3 (long pauses after a hint), with a cutoff of 8 s for "long"; Feature 4 (long pauses after a hint and correct answer), with a cutoff of 12 s for "long"; Feature

6 (off-task behavior); Feature 7 (gaming the system); Feature 7' (fast non-gaming actions), with a cutoff of 2 s for "fast"; and Feature 9' (spikiness in moment-by-moment learning).

Seven out of nine of these features depend on a threshold parameter, N ; adjusting a feature's parameter can result in a very different model. For each of these features, we used brute-force grid search to find an optimal cutoff level for each of the above-mentioned features (in grid search, values are tried for every step at the same interval—for instance, 0.5 s, 1 s, 1.5 s, 2 s, etc.). Optimality was defined in terms of the ability to predict the dependent variable, performance on the transfer test. Variables involving probabilities were searched at a grid size of 0.05; variables involving time were searched at a grid size of 0.5 s.

The cross-validated correlations for single-feature regression models are shown in Table 1.

These nine features were considered as potential candidates for a unified model (other features, which individually had cross-validated correlations below zero, were eliminated from consideration, as a control on overfitting). To find a unified model combining multiple parameters, Forward Selection was conducted (Ramsey & Schafer, 1997). In Forward Selection, the best single-parameter model is chosen, and then the parameter that most improves the model is repeatedly added until no more parameters can be added that improve the model. The goodness metric used was the LOOCV correlation between the predictions and each student's performance on the transfer test.

The resultant model was

$$\text{Transfer} = -1.5613 \times \text{HelpAvoidance} (1) \\ + 0.2968 \times \text{FastNotGaming} (7') + 0.8272.$$

The feature most strongly associated with transfer, both by itself and as a component of a unified model, was avoiding help, which was negatively associated with transfer (cross-validated $r = .376$). One potential interpretation is that help avoidance directly caused lower learning (cf. Aleven et al., 2006), perhaps causing the students to have less conceptual learning, as the tutor hints are fairly conceptual in nature. This lack of conceptual understanding may in turn have made these students less able to transfer their knowledge. The other individual feature incorporated into the model was fast nongaming actions. These actions were significantly positively associated with transfer. Fast nongaming actions

Table 1
Goodness of Single-Feature Linear Regression Models for Predicting Transfer in the College Data Set

Feature	Transfer =	Cross-validated r
1. Help avoidance	$-1.735 \times F1 + 0.912$.376
9'. Spikiness of moment-by-moment learning	$-9.758 \times F9 + 0.951$.346
4. Long pauses after reading hint messages and then getting the next action right	$-6.510 \times F4 + 0.893$.204
3. Long pauses after reading hint messages	$-4.075 \times F3 + 0.902$.199
7'. Fast actions that do not involve gaming	$0.484 \times F7' + 0.726$.188
2. Long pauses after receiving bug messages	$-13.497 \times F2 + 0.880$.130
7. Gaming the system	$-0.2058 \times F7 + 0.903$.076
2'. Short pauses after receiving bug messages	$-4.291 \times F2' + 0.876$.037
5. Off-task behavior	$-1.037 \times F5 + 0.899$.024

may indicate a degree of fluency with the relevant skills that facilitates reasoning with them, as hypothesized by Haverty, Koedinger, Klahr, and Alibali (2000), leading to better transfer.

The cross-validated correlation of the model to the transfer test was .396, as shown in Table 2.

PFL Detector

The same set of 18 features and model development process described in the previous section was used to develop a model of students' PFL. In the case of PFL, five features showed positive cross-validated correlations between the individual feature and the students' performance on the PFL test: Feature 1 (help avoidance), with a cutoff of 85% probability for "poorly known"; Feature 3 (long pauses after a hint), with a cutoff of 8 s for "long"; Feature 3' (short pauses after a hint), with a cutoff of 1 s for "short"; Feature 4 (long pauses after a hint and correct answer), with a cutoff of 8 s for "long"; Feature 4' (short pauses after a hint and correct answer), with a cutoff of 20 s for "short"; Feature 6 (off-task behavior); Feature 6' (long pauses that are not off-task), with a cutoff of 4 s for "long"; Feature 7 (gaming the system), Feature 7' (fast non-gaming actions), with a cutoff of 4 s for "fast"; Feature 9 (average moment-by-moment learning); and Feature 9' (spikiness in moment-by-moment learning).

Single-feature regression models fit on the whole data set, and their associated cross-validated correlations are shown in Table 3 (only features with cross-validated correlation over zero are shown).

These 11 features were considered as potential candidates for a unified model. To find a unified model combining multiple parameters, Forward Selection was conducted, as with the transfer model.

The resultant models was

$$\begin{aligned} \text{PFL} = & 0.0127 \times \text{Spikiness (9)} - 0.5499 \times \text{HelpAvoidance (1)} \\ & - 5.3898 \times \text{LongPauseAfterHint (4)} + 0.8773. \end{aligned}$$

The feature most strongly associated with PFL was long pauses after reading hint messages and getting the next action correct, which was somewhat unexpectedly negatively associated with PFL (cross-validated $r = .410$). As with transfer, help avoidance was also negatively associated with PFL (cross-validated $r = .329$), and entered into the final model. Finally, the spikiness of the student's learning is positively associated with PFL, and enters into the final model, achieving a cross-validated r of .233. This finding suggests that PFL is higher if a student's learning more frequently occurs in relatively sudden "aha" moments, as compared with occurring more gradually; deeper learning is occurring.

Table 2
Cross-Validated Correlations Between Models and Tests

Construct	Data developed with	Data tested on	Cross-validated r
Transfer	College	College	.396
Transfer	College	High school	.426
Transfer	High school	High school	.528
PFL	College	College	.454
PFL	College	High school	.228
PFL	High school	High school	.181

Note. PFL = preparation for future learning.

As shown in Table 2, the overall cross-validated correlation of the model to the PFL test was .454.

Transfer and PFL

Given the existence of models that can predict PFL and transfer to a reasonable degree, one question is to what degree these two models are capturing the same construct. The two constructs have a fairly substantial correlation of .520. However, it is worth studying whether the two forms of robust learning are characterized by the same behaviors during learning.

The results of these two models seem to suggest substantial overlap. First, several of the same data features were found to be associated with both transfer and PFL under cross-validation: 1, 3, 4, 5, 7, 7', and 9'. In fact, only two features predicted transfer but failed to predict PFL, and only four features predicted PFL but failed to predict transfer.

In addition, each model was successful at predicting the other construct. When used to predict PFL, the optimized-feature transfer detector achieves a correlation of .425, almost as good as the optimized model trained to predict PFL. Correspondingly, when used to predict transfer, the optimized-feature PFL detector achieves a correlation of .395, almost identical to the detector trained just to predict transfer.

Studying the Goodness-of-Transfer and PFL Detectors for High School Data

After developing these detectors, our next goal was to understand how well these detectors transfer between different populations of students. To this end, data were analyzed for a sample of high school students working with the same Genetics Cognitive Tutor module to examine whether the robust learning models transfer between two populations who vary in age and prior preparation.

Data Set

As in the original study, the data used in the second study came from the Genetics Cognitive Tutor Three-Factor Cross module. Fifty-six high school students who were enrolled in high school biology courses used the tutor. The students were recruited to participate in the study for pay through several methods, including advertisements in a regional newspaper and recruitment handouts distributed at two urban high schools.

The study had the same design as the college-level study. In specific, it consisted of two 2-hr sessions, followed by a shorter session 1 week later, all conducted in computer clusters at Carnegie Mellon University. The students engaged in Cognitive Tutor-supported activities for 1 hr in each of two sessions. As in the original study, students completed a transfer test and PFL test after using the tutor, as well as completing a pretest and posttest of the exact skills taught in the tutor. All tests were identical to the ones used in the previous study.

The 56 students completed a total of 21,498 problem-solving attempts across a total of 9,204 problem steps in the tutor. The number of problem-solving attempts per student was not significantly different between the college and high school populations, $t(126) = 0.847$, $p = .40$. Like the college students, the high school students demonstrated successful learning in this tutor, with an

Table 3
Goodness of Single-Feature Linear Regression Models for Predicting PFL in the College Data Set

Feature	PFL =	Cross-validated <i>r</i>
4. Long pauses after reading hint message(s) and then getting the next action right	$-7.67 \times F4 + 0.961$.410
3. Long pauses after reading hint messages	$-5.050 \times F3 + 0.956$.376
9. Average moment-by-moment learning	$-8.240 \times F9 + 0.979$.345
1. Help avoidance	$-1.118 \times F1 + 0.952$.329
9'. Spikiness of moment-by-moment learning	$0.022 \times F9 + 0.740$.233
4'. Short pauses after reading hint message(s) and then getting the next action right	$-1.801 \times F4' + 0.937$.201
7'. Fast actions that do not involve gaming	$0.350 \times F7' + 0.739$.187
5. Off-task behavior	$-1.089 \times F5 + 0.944$.089
5'. Long pauses that are not off-task	$-0.211 \times F5' + 0.976$.083
3'. Short pauses after reading hint messages	$0.173 \times F3' + 0.886$.034
7. Gaming the system	$-0.134 \times F7 + 0.93$.008

Note. PFL = preparation for future learning.

average pretest performance of 0.16 ($SD = 0.09$) and an average posttest performance of 0.56 ($SD = 0.28$), a statistically significant difference, $t(55) = 11.443$, $p < .001$. Students' average transfer test performance was 0.53 ($SD = 0.22$) and average PFL performance was 0.66 ($SD = 0.28$).

Transferring Robust Learning Detectors From College Students to High School Students

To check the generalizability of the transfer and PFL detectors, we tested the predictive power of each detector, taking the detectors developed and optimized using the college data and applying them without modification to the high school data set.

The college detector of transfer achieved a correlation of .426 to the transfer test scores within the high school data set. It is worth noting that this correlation was higher than the correlation (.396) in the college data set, despite the model being transferred to a new population. One possible explanation is that there is a closer link between in-tutor performance and transfer test performance in the high school population than the college population, potentially because students were closer to reaching the performance ceiling in the original college population.

By contrast, the college detector of PFL achieved a correlation of .228 to the PFL test scores within the high school data set, a value that represents substantial degradation compared with the data set for which these models was originally developed (where the value was 0.454). At the same time, this model remains marginally statistically significantly higher than zero ($p = .09$).

Building New Robust Learning Detectors for High School Students

In order to fully understand the degree of degradation between the college and high school populations, we can build new detectors for the high school population. Seeing how well these detectors perform can give us an upper limit for how well this type of detector can perform in this data set. It also may be interesting to study which data features are important predictors within the high school population, to see how these features differ from those used in the college population, at a qualitative level.

A new detector of transfer trained on the data from the high school population using optimized features achieves a cross-validated correlation of .528. This number is moderately higher than the goodness of the detector trained on the college population and then applied to this data set, which was .426. It is also higher than the performance of the goodness of the detector trained on the college population on its original data set, which was .396, again indicating that student behavior is more closely linked to performance on the transfer test in the high school population than in the college population.

By contrast, a new detector of PFL trained on the data from the high school population using optimized features achieves an unimpressive cross-validated correlation of .181. This number is actually lower than the goodness of the detector trained on the college population and then applied to this data set, which was .228. It is also substantially lower than the performance of the goodness of the detector trained on the college population on its original data set, which was .454. This result indicates that the behaviors associated with PFL in this new population are not captured well by the feature set originally developed within the college population.

Features Associated With Robust Learning in High School Data Set: Transfer

Within the high school data set, 13 individual features were found to have positive cross-validated correlation to the transfer test scores. The single-feature linear regression model for each feature is given in Table 4.

There was substantial overlap between the features that had positive cross-validated correlations in the college and high school populations. Only one of the features that had a positive cross-validated correlation for the college population failed to have a positive cross-validated correlation for the high school population, short pauses after bug messages (Feature 2). Of the remaining features, all but two pointed in the same direction in both data sets (pointing in the same direction means that the model coefficient was either negative in both data sets or positive in both data sets). The two that changed direction were the spikiness of moment-by-moment learning (negative in the college data set and positive in the high school data set) and off-task behavior (negative in the college data set and positive in the high school data set). It is worth noting that off-task behavior had the weakest relationship that still had a positive cross-

Table 4
Goodness of Optimized Single-Feature Linear Regression Models at Predicting Transfer in High School Data Set

Feature	Transfer =	Cross-validated <i>r</i>
7. Gaming the system	$-0.9108 \times F7 + 0.8482$.496
9. Average moment-by-moment learning	$-16.6448 \times F9 + 0.906$.490
7'. Fast actions that do not involve gaming	$0.8805 \times F7' + 0.0374$.437
8. Average contextual slip	$1.4064 \times F8 + 0.0226$.429
8'. Certainty of slip	$0.8412 \times F8 + 0.2947$.409
3'. Short pauses after reading hint messages	$-1.2538 \times F3' + 0.6355$.396
3. Long pauses after reading hint messages	$-1.3839 \times F3 + 0.6512$.391
1. Help avoidance	$-1.6946 \times F1 + 0.7475$.386
4. Long pauses after reading hint message(s) and then getting the next action right	$-1.5936 \times F4 + 0.6321$.367
9'. Spikiness of moment-by-moment learning	$0.0598 \times F9 + 0.2722$.362
4'. Short pauses after reading hint message(s) and then getting the next action right	$-1.3071 \times F4' + 0.61$.350
2. Long pauses after bug messages	$-43.8096 \times F2 + 0.5588$.200
5. Off-task behavior	$1.7228 \times F5 + 0.4554$.051

validated correlation, in both data sets (.024 and .051). Hence, the primary noteworthy difference is the relationship for spikiness.

That said, it is worth noting that many of the features changed semantics substantially during parameter optimization. Only one feature retained similar semantics between the two data sets, help avoidance (Feature 1), which had an optimized cutoff of 70% in the college data set, but an optimized cutoff of 50% in the high school data set, a relatively minor change. In terms of features that changed semantics, Feature 3, long pauses after reading help messages, changed from a cutoff of 8 s in the college data set to 1 s in the high school data set, a substantially different feature. Similarly, Feature 4, long pauses after reading help messages and then obtaining a correct answer, changed from 12 s in the college data set to 1 s in the high school data set. Feature 7', fast non-gaming actions, shifted in the other direction, from 2 s to 20 s.

Five additional features were also significant in the high school model: Feature 3' (short pauses after a hint), with a cutoff of 17 s for "short"; Feature 4' (short pauses after a hint and correct answer), with a cutoff of 17 s for "short"; Feature 8 (average contextual slip); Feature 8' (certainty of contextual slip); Feature 9 (average moment-by-moment learning).

A model was fit using Forward Selection, as in the college data set. The best model of transfer for the high school data set, using the optimal feature cutoffs, and fitting to all data, was as follows:

$$\text{Transfer} = -0.793 \times \text{Gaming (7)} + 1.518$$

$$\times \text{Off-task behavior (6)} - 34.429$$

$$\times \text{LongPauseAfterBug (2)} + 0.7587.$$

Features Associated With Robust Learning in High School Data Set: PFL

A range of variables were found to have cross-validated correlations over zero to the PFL test within the high school population, shown in Table 5. There was considerable overlap between the college and high school populations for these features. Seven of the 11 features used in the college detector of PFL were also used in the high school detector of PFL (Feature 3, Feature 3', Feature 4, Feature 7, Feature 7', Feature 9, Feature 9'), with all pointing in the same direction in the two data sets except for Feature 3', which switched direction.

However, none of these features had particularly impressive correlations taken individually, with the highest cross-validated correlation for the high school data set having a value of .137. This feature was Feature 9', the spikiness of the moment-by-moment learning model. Two other features had cross-validated correlations of .1 or higher: the certainty of slip and gaming the system. Spikiness and gaming were also found in the college PFL model,

Table 5
Goodness of Optimized Single-Feature Linear Regression Models at Predicting PFL in High School Data Set

Feature	PFL =	Cross-validated <i>r</i>
9'. Spikiness of moment-by-moment learning	$0.045 \times F9' + 0.4622$.137
8'. Certainty of slip	$0.5802 \times F8' + 0.4941$.123
7. Gaming the system	$-0.5002 \times F7 + 0.8316$.105
3. Long pauses after reading hint messages	$-1.637 \times F3 + 0.752$.097
9. Average moment-by-moment learning	$-9.195 \times F9 + 0.865$.092
4. Long pauses after reading hint message(s) and then getting the next action right	$-2.3075 \times F4 + 0.7452$.073
2. Long pauses after bug messages	$-30.6071 \times F2 + 0.6819$.059
3'. Short pauses after reading hint messages	$-0.744 \times F3' + 0.7193$.049
8. Average contextual slip	$0.7828 \times F8 + 0.3744$.045
7'. Fast actions that do not involve gaming	$0.4773 \times F7' + 0.3899$.041

Note. PFL = preparation for future learning.

where the relationships pointed in the same direction as in the high school data set.

A model of PFL was fit using Forward Selection, as in the college data set. The best model of PFL for the high school data set, using the optimal feature cutoffs, and fitting to all data, was as follows:

$$\begin{aligned} \text{PFL} = & 0.028 \times \text{Spikiness (9')} \\ & - 1.1901 \times \text{LongPauseAfterHint (3)} \\ & - 27.343 \times \text{LongPauseAfterBug (2)} + 0.6214. \end{aligned}$$

Conclusions

In this article, we have studied the degree to which automated detectors of transfer and PFL transfer to a new cohort of students, using the same tutor lesson. These findings establish that it is not just possible to identify whether a student has achieved robust learning; it is also possible to successfully apply these models on a different population than the initial population these detectors were developed for, establishing that there is some degree of generality in the constructs that these detectors tap.

The detector of transfer generalized from the college population to the high school population with limited evidence of degradation; in fact, the detector functioned better within the new population than in the original population, though not quite as well as a new detector trained specifically for the new population.

The detector of PFL, however, saw relatively greater evidence of degradation between the college and high school population, achieving a correlation only about half as high within the high school population as had been achieved within the college population. However, it may just be that PFL was relatively difficult to detect within the high school population, as a detector trained specifically for the new population also functioned relatively poorly.

Between the high school and college populations, many of the same features were predictive of transfer and PFL. There was substantial overlap in both cases, with seven of nine features that had cross-validated correlation over zero in the college data set achieving a cross-validated correlation over zero and a coefficient pointing in the same direction as in the college model, when transferred to the high school data set. Six of 11 features achieved this same standard when the college model of PFL was transferred to the high school data set, a lower degree of overlap but still an indication of considerable similarity between the construct in the two data sets.

Four features were predictive (and pointed in the same direction) in every model: Features 3, 4, 7, and 7'. Feature 3, long pauses after reading hint messages, and Feature 4, long pauses after reading hint messages and providing a correct answer, were negatively correlated with robust learning for each construct and data set. This does not necessarily mean that these pauses (interpreted as implying self-explanation; cf. Shih et al., 2008) actually hurt learning, but may instead indicate a general selection bias where the students who seek help are generally less knowledgeable (cf. Aleven et al., 2006). These results build on past findings regarding relationships between students' strategies for using help and their learning outcomes (cf. Aleven et al., 2006). We recom-

mend that future research on help seeking and learning consider measures of transfer and PFL to a greater degree.

Feature 7, gaming the system, was also negatively correlated with robust learning for each construct and data set, albeit with relatively low correlations. This finding accords with previous results suggesting that gaming the system is particularly pernicious for learning (cf. Cocea, Hershkovitz, & Baker, 2009).

However, fast nongaming actions were positively correlated with robust learning for each construct and data set, with generally strong correlations. These actions appear to indicate robust learning that leads to both transfer and PFL. Given that fast correct actions are also associated with retention (cf. Pavlik & Anderson, 2008), it appears that rapid correct performance indicates learning that is robust in multiple fashions.

Many other features were associated with robust learning for a single construct. Help avoidance was associated with transfer with a strong negative correlation in both populations. Previous analysis has also revealed negative correlations between help avoidance and learning; for example, students who make errors when they should have sought help perform more poorly on tests of standard problem solving (Aleven et al., 2006). Help in the Genetics Cognitive Tutor is fairly conceptual in nature; that is, it relates the steps in the problem-solving procedure to the properties of the underlying genetic processes. Our findings suggest that this type of help is associated not just with learning to solve the types of problems in the tutor, but leads to robust learning as well. Prior work studying the learning impact of teaching students when to seek help has not had significant effects on problem-solving posttests (Roll, Aleven, McLaren, & Koedinger, 2011); it would be worth studying whether this type of meta-cognitive instruction impacts performance on measures of robust learning, even if it does not impact performance on problem-solving posttests. An alternate explanation for the negative relationship between help avoidance and robust learning in our study—in line with the results in Roll et al. (2011)—is that some students are not prepared to learn from the types of help in the tutor, leading them to both avoid help and demonstrate less robust learning. In general, further attention to why students avoid help and how students use help successfully and unsuccessfully (cf. Aleven, Stahl, Schworm, Fischer, & Wallace, 2003) may help us better understand this finding.

Features of the moment-by-moment learning model were associated with PFL in both data sets. In particular, spikiness as measured by the moment-by-moment learning model was positively associated with PFL in both data sets. In other recent work, researchers have suggested that further distillations of the moment-by-moment graph, in particular through explicitly considering the visual form of the graph, can be even more predictive of student PFL (Baker, Hershkovitz, Rossi, Goldstein, & Gowda, in press).

In general, this article suggests that models of robust learning can be transferred to new populations. As such, these models can be used with relative confidence for new groups of students, to drive interventions. By doing so, we can move toward the vision of learning systems that can adapt effectively to individual differences not just in what students know, but in how robust their learning is.

Another valuable area of future work will be to determine how general the phenomena seen here are for new content: new lessons within the Genetics Tutor, Cognitive Tutors on other topics, and additional learning systems. The models presented here are time-consuming in nature to develop; to the extent that general models can

be developed, their potential usefulness will be substantially increased.

References

- Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26, 147–179.
- Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence and Education*, 16, 101–128.
- Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. M. (2003). Help seeking and help design in interactive learning environments. *Review of Educational Research*, 73, 277–320.
- Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, 94, 416–427.
- Bahrack, H. P., Bahrack, L. E., Bahrack, A. S., & Bahrack, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, 4, 316–321.
- Baker, R. S. J. d. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1059–1068). New York, NY: Association for Computing Machinery.
- Baker, R. S. J. d., Corbett, A. T., Gowda, S. M., Wagner, A. Z., McLaren, B. M., Kauffman, L. R., . . . Giguere, S. (2010). Contextual slip and prediction of student performance after use of an intelligent tutor. In P. De Bra, A. Kobsa, & D. Chin (Eds.), *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization* (pp. 52–63). Heidelberg, Germany: Springer-Verlag.
- Baker, R. S. J. d., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, 18, 287–314.
- Baker, R. S. J. d., Gowda, S. M., & Corbett, A. T. (2011a). Automatically detecting a student's preparation for future learning: Help use is key. In *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 179–188). Retrieved from <http://alexandria.tue.nl/repository/books/715601.pdf>
- Baker, R. S. J. d., Gowda, S., & Corbett, A. T. (2011b). Towards predicting future transfer of learning. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 23–30). Heidelberg, Germany: Springer-Verlag.
- Baker, R. S. J. d., HersHKovitz, A., Rossi, L. M., Goldstein, A. B., & Gowda, S. M. (in press). Predicting robust learning with the visual form of the moment-by-moment learning curve. *Journal of the Learning Sciences*.
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24, 61–100.
- Butcher, K. R. (2010). How diagram interaction supports learning: Evidence from think alouds during intelligent tutoring. *Lecture Notes in Computer Science*, 6170, 295–297.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182.
- Chi, M. T. H., Feltoovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Chi, M., & VanLehn, K. (2007). Domain-specific and domain-independent interactive behaviors in Andes. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 548–550). Amsterdam, the Netherlands: IOS Press.
- Chin, D. B., Dohmen, I. M., Cheng, B. H., Oppezzo, M. A., Chase, C. C., & Schwartz, D. L. (2010). Preparing students for future learning with teachable agents. *Educational Technology Research and Development*, 58, 649–669.
- Cocca, M., HersHKovitz, A., & Baker, R. S. J. d. (2009). The impact of off-task and gaming behaviors on learning: Immediate or aggregate? In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. C. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 507–514). Amsterdam, the Netherlands: IOS Press.
- Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, 12, 371–417.
- Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. Gmytrasiewicz, & J. Vassileva (Eds.), *UM2001, User modeling: Proceedings of the Eighth International Conference* (pp. 137–147). Berlin, Germany: Springer.
- Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4, 253–278.
- Corbett, A. T., & Bhatnagar, A. (1997). Student modeling in the ACT Programming Tutor: Adjusting procedural learning model with declarative knowledge. In A. Jameson, C. Paris, & C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference* (pp. 243–254). New York, NY: Springer.
- Corbett, A. T., McLaren, B., Kauffman, L., Wagner, A., & Jones, E. (2010). A Cognitive Tutor for genetics problem solving: Learning gains and student modeling. *Journal of Educational Computing Research*, 42, 219–239.
- Corbett, A., McLaren, B., Wagner, A., Kauffman, L., Mitchell, A., Baker, R., & Gowda, S. (2011). Preparing students for effective explaining of worked examples in the Genetics Cognitive Tutor. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the Thirty-Third Annual Meeting of the Cognitive Science Society* (pp. 1476–1481). Wheat Ridge, CO: Cognitive Science Society.
- Corbett, A. T., & Trask, H. (2000). Instructional interventions in computer-based tutoring: Differential impact on learning time and accuracy. In *Proceedings of the ACM CHI '2000 Conference on Human Factors in Computing Systems* (pp. 97–104). Reading, MA: ACM Press.
- Craig, S., VanLehn, K., & Chi, M. (2008). Promoting learning by observing deep-level reasoning questions on quantitative physics problem solving with Andes. In K. McFerrin, R. Weber, R. Carlsen, & D. A. Willis (Eds.), *Proceedings of the Society for Information Technology and Teacher Education International Conference 2008* (pp. 1065–1068). Chesapeake, VA: AACE.
- Desmarais, M. C., Meshkinfam, P., & Gagnon, M. (2006). Learned student models with item to item knowledge structures. *User Modeling and User-Adapted Interaction*, 16, 403–434.
- Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, 37, 36–48.
- Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General*, 120, 34–45.
- Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of learning: Contemporary research and applications. The educational technology series* (pp. 9–46). San Diego, CA: Academic Press.
- Gong, Y., Beck, J. E., & Heffernan, N. T. (2010). Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 35–44). Heidelberg, Germany: Springer-Verlag.
- Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue

- in natural language. *Behavioral Research Methods, Instruments, & Computers*, 36, 180–192.
- Hausmann, R., & VanLehn, K. (2007). Explaining self-explaining: A contrast between content and generation. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic, (Eds.), *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (pp. 417–424). Heidelberg, Germany: Springer-Verlag.
- Haverty, L. A., Koedinger, K. R., Klahr, D., & Alibali, M. W. (2000). Solving inductive reasoning problems in mathematics: Not-so-trivial pursuit. *Cognitive Science*, 24, 249–298.
- Jastrzemski, T. S., Gluck, K. A., & Gunzelmann, G. (2006). Knowledge tracing and prediction of future trainee performance. In *Proceedings of the 2006 Interservice/Industry Training, Simulation, and Education Conference* (pp. 1498–1508). Orlando, FL: National Training Systems Association.
- Katz, S., Connelly, J., & Wilson, C. (2007). Out of the lab and into the classroom: An evaluation of reflective dialogue in ANDES. In *Proceedings of the 2007 conference on Artificial Intelligence in Education* (pp. 425–432). Amsterdam, the Netherlands: IOS Press.
- Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–77). New York, NY: Cambridge University Press.
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction (KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36, 757–798.
- Li, N., Cohen, W. W., & Koedinger, K. R. (2010). A computational model of accelerated future learning through feature recognition. In *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 368–370). Berlin, Germany: Springer.
- Martin, J., & VanLehn, K. (1995). Student assessment using Bayesian nets. *International Journal of Human-Computer Studies*, 42, 575–591.
- Mathan, S. A., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational Psychologist*, 40, 257–265.
- McCalla, G. (1992). The search for adaptability, flexibility and individualization: Approaches to curriculum in ITS. In M. Jones & P. Winne (Eds.), *Adaptive learning environments: Foundations and frontiers* (pp. 91–122). Berlin, Germany: Springer-Verlag.
- McLaren, B. M., Lim, S., & Koedinger, K. R. (2008). When and how often should worked examples be given to students? New results and a summary of the current state of research. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2176–2181). Austin, TX: Cognitive Science Society.
- Pardos, Z. A., Gowda, S. M., Baker, R. S. J. d., & Heffernan, N. T. (2011). Ensembling predictions of student post-test scores for an intelligent tutoring system. In *Proceedings of the 4th International Conference on Educational Data Mining* (pp. 189–198). Retrieved from <http://alexandria.tue.nl/repository/books/715601.pdf>
- Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14, 101–117.
- Pavlik, P. I., Cen, H., & Koedinger, J. R. (2009). Performance factors analysis—A new alternative to knowledge tracing. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. C. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 531–540). Amsterdam, the Netherlands: IOS Press.
- Pavlik, P. I., Cen, H., Wu, L., & Koedinger, K. R. (2008). Using item-type performance covariance to improve the skill model of an existing tutor. In R. S. Baker & J. E. Beck (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 77–86). Worcester, MA: International Educational Data Mining Society.
- Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology*, 25, 111–195.
- Ramsey, R. L., & Schafer, D. W. (1997). *The statistical sleuth*. Belmont, CA: Wadsworth.
- Rittle-Johnson, B., & Siegler, R. S. (1998). The relation between conceptual and procedural knowledge in learning mathematics: A review. In C. Donlan (Ed.), *The development of mathematical skill* (pp. 75–110). Hove, UK: Psychology Press.
- Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2007). Designing for metacognition—Applying cognitive tutor principles to the tutoring of help seeking. *Metacognition and Learning*, 2, 125–140.
- Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21, 267–280.
- Salden, R. J. C. M., Koedinger, K. R., Renkl, A., Aleven, V., & McLaren, B. M. (2010). Accounting for beneficial effects of worked examples in tutored problem solving. *Educational Psychology Review*, 22, 379–392.
- Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3, 207–217.
- Schofield, J. W. (1995). *Computers and classroom culture*. Cambridge, UK: Cambridge University Press.
- Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, 22, 129–184.
- Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior*, 25, 258–266.
- Shih, B., Koedinger, K. R., & Scheines, R. (2008). A response time model for bottom-out hints as worked examples. In R. S. J. d. Baker & J. E. Beck (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 117–126). Worcester, MA: International Educational Data Mining Society.
- Shute, V. J. (1995). SMART: Student modeling approach for responsive tutoring. *User Modeling and User-Adapted Interaction*, 5, 1–44.
- Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.
- Tan, J., & Biswas, G. (2006). The role of feedback in preparation for future learning: A case study in learning by teaching environments. In M. Ikeda, K. Ashley, & T.-W. Chan (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 370–381). Heidelberg, Germany: Springer-Verlag.
- VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16, 227–265.
- VanLehn, K., Jones, R., & Chi, M. T. H. (1992). A model of the self-explanation effect. *Journal of the Learning Sciences*, 2, 1–59.
- Wang, Y., & Heffernan, N. (2011). The “assistance” model: Leveraging how many hints and attempts a student needs. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International FLAIRS Conference*. Palo Alto, CA: AAAI Press.

Received December 9, 2011

Revision received March 5, 2013

Accepted April 5, 2013 ■

Gender Differences in the Use and Benefit of Advanced Learning Technologies for Mathematics

Ivon Arroyo
University of Massachusetts Amherst

Winslow Burleson
Arizona State University

Minghui Tai
University of Massachusetts Amherst

Kasia Muldner
Arizona State University

Beverly Park Woolf
University of Massachusetts Amherst

We provide evidence of persistent gender effects for students using advanced adaptive technology while learning mathematics. This technology improves each gender's learning and affective predispositions toward mathematics, but specific features in the software help either female or male students. Gender differences were seen in the students' style of use of the system, motivational goals, affective needs, and cognitive/affective benefits, as well as the impact of affective interventions involving pedagogical agents. We describe 4 studies, with hundreds of students in public schools over several years, which suggest that technology responses should probably be customized to each gender. This article shows differential results before, during, and after the use of adaptive tutoring software, indicating that digital tutoring systems can be an important supplement to mathematics classrooms but that male and female students should be addressed differently. Female students were more receptive than male students to seeking and accepting help provided by the tutoring system and to spending time seeing the hints; thus, they had a consistent general trend to benefit more from it, especially when affective learning companions were present. In addition, female students expressed positively valenced emotions most often and exhibited more productive behaviors when exposed to female characters; these affective pedagogical agents encouraged effort and perseverance. This was not the case for male students, who had more positive outcomes when no learning companion was present and their worst affective and cognitive outcomes when the female character was present.

Keywords: adaptive learning environments, gender differences, affective agents, motivation and affect, quantitative analysis

Advanced learning technologies have the potential to personalize instruction for students and to meet individual learning needs. To provide such personalized instruction, researchers must first assess factors that influence student learning. A key factor in the domain of mathematics is student gender. To date, much of the educational psychology research on motivation and achievement has been conducted in standard classrooms with-

out technology support. This article provides an analysis of the impact of gender in mathematics learning with instructional software, called *Wayang Outpost*. Gender differences were investigated by focusing on cognitive and affective factors in learning. Findings from this research form a theoretical foundation for the design of adaptive computational tutors that deliver personalized support.

This article was published Online First September 16, 2013.

Ivon Arroyo, Department of Computer Science, University of Massachusetts Amherst; Winslow Burleson, Department of Computing, Informatics, and Decision Systems Engineering, Arizona State University; Minghui Tai, Department of Teacher Education and Curriculum Studies, University of Massachusetts Amherst; Kasia Muldner, Department of Computing, Informatics, and Decision Systems Engineering, Arizona State University; Beverly Park Woolf, Department of Computer Science, University of Massachusetts Amherst.

We gratefully acknowledge support for this work from National Science Foundation Grants HRD/EHR 1109642 ("Personalized Learning: Strategies to Respond to Distress and Promote Success") to I. Arroyo, B. P. Woolf, and W. Burleson; HRD/EHR 012080 ("AnimalWorld: Enhancing High School Women's Mathematical Competence") to C. Beal, B. P. Woolf, and J. M. Royer; HRD/EHR

0411776 ("Learning to Teach: The Next Generation of Intelligent Tutor Systems") to B. P. Woolf, I. Arroyo, A. Barto, S. Mahdevan, and D. Fisher; HRD GSE/RES 0734060 ("What Kind of Math Software works for Girls?") to I. Arroyo, J. M. Royer, and B. P. Woolf; and IIS/HCC 0705554 ("Affective Learning Companions: Modeling and supporting emotion during teaching") to B. P. Woolf, W. Burleson, I. Arroyo, A. Barto, and D. Fisher, as well as a "Teaching Every Student" grant from the U.S. Department of Education to B. P. Woolf, I. Arroyo, and R. Maloy. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agencies.

Correspondence concerning this article should be addressed to Ivon Arroyo, Department of Computer Science, 140 Governors Drive, University of Massachusetts Amherst, Amherst, MA 01003-9264. E-mail: ivon@cs.umass.edu

Literature Review of Gender Differences and Performance

A wealth of motivational and achievement educational psychology research has established the development of gender differences in mathematics education, both in the affective and the cognitive domain. This section summarizes the literature regarding these two aspects, affective and cognitive.

In relation to achievement, girls generally tend to have higher grades in math classes, but boys tend to score higher than girls in standardized tests (Hyde, Lindberg, Linn, Ellis, & Williams, 2008). Female students tend to enroll in less advanced math classes, although this trend has been reducing. Various cognitive factors have been held responsible for the difference in performance in math tests such as SAT Math and National Assessment of Educational Progress. Among them are gender differences in basic spatial abilities such as mental rotations (Casey, Nuttall, Pezaris, & Benbow, 1995), which are heavily involved not only in highly visual (e.g., geometry) math problems but also in mathematics problems that require approximate solutions and estimations of magnitude. Differences in verbal components such as the speed and accuracy of retrieval of basic arithmetic facts from long-term memory into working memory (Royer, Tronsky, Chan, Jackson, & Marchant, 1999) are another possibility,¹ because phonological areas of the brain are heavily involved in exact solutions to mathematics problems and arithmetic (Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999). Although both nature and nurture may be at play to contribute to these differences, evidence has suggested that nurture components are heavily at play. For instance, studies have shown that action video games can virtually eliminate a gender difference in spatial attention (in only 10 hours of play) and significantly decrease the gender disparity in mental rotation ability (Feng, Spence, & Pratt, 2007). Similarly, studies have shown how basic math facts retrieval can be trained for both speed and accuracy and that this in turn improves mathematics problem-solving ability (Arroyo, Royer, & Woolf, 2011; Royer et al., 1999). Thus, a clear possibility is that female and male students are not exposed to activities that develop these two core areas related to mathematics cognition in equal ways, as children grow up, either inside or outside of school.

The affective component of gender differences in mathematics is related to the fact that, as they progress throughout the K–12 school system, girls report increasingly more negative attitudes toward mathematics and express more self-derogating attributions about their mathematics performance (Hyde et al., 2008; Royer & Walles, 2007). In particular, gender differences have been found in early adolescence for mathematics self-concept (belief about one's ability to learn mathematics) and mathematics utility (belief that mathematics is important; Eccles, Wigfield, Harold, & Blumenfeld, 1993). Additionally, both female and minority students develop more negative feelings toward mathematics during their school years than do the rest of students (Catsambis, 2005), although there are important racial-ethnic differences, with affective gender differences being most pronounced among Latino students.

It is possible that these affective differences are either a cause or a consequence of the gender difference in mathematics performance in standardized tests. Stereotype threat, or the concern that others will view one stereotypically (Spencer, Steele, & Quinn, 1999), has been identified to account for gender differences in

mathematical problem solving, suggesting that female students might receive messages—from peers, parents, teachers, or the media—about a possible performance superiority for male math ability or about female inferiority.

However, it is important to note that one of the largest studies (Catsambis, 2005) involving thousands of students showed there is a trend for all students to decrease their interest in mathematics as they progress throughout the school system and increase their perception of the difficulty of mathematics, in contrast to other school subjects. Thus, generating positive experiences in mathematics learning for all students, but for female and minority students in particular, should be an important goal of mathematics education research and practice.

The Impact of School and Educational Practices on Gender Development

The school functions as a primary setting for developing gender orientations. Studies over several past decades have indicated that teachers used to pay more attention to boys and interacted with them more extensively (Ebbeck, 1984; Fennema, Carpenter, Jacobs, Franke, & Levi, 1998; Forgasz & Leder, 2006). Boys used to receive more praise as well as criticism from teachers in the classroom than did girls (Cherry, 1975), and they were more likely to be praised for academic success and criticized for misbehavior. Girls tended to be praised for tidiness and compliance and criticized for academic failure, which could undermine their perceived self-efficacy (Eccles, 1987). On the other hand, when teachers emphasized the usefulness of quantitative skills and encouraged cooperative or individualized rather than competitive learning, female students showed higher perceived efficacy and valuation of mathematics (Eccles, 1989).

Gender differences have been observed in how male versus female students judge their math capabilities with impact on math course selection (Eccles, 1987; Hyde & Linn, 1988). Until some years ago, female students were known to enroll in significantly fewer higher level mathematics, science, and computer courses; to have less interest in these subjects; and to view these course as less useful than did their male counterparts. For decades, school counselors have encouraged and supported the interest of male students in scientific fields and have steered female students away from scientific and technical fields (Fitzgerald & Crites, 1980). Many efforts have been made in the past decades to change such practices (National Council of Teachers of Mathematics, 2008; Sevo & Chubin, 2010). Meanwhile, cross-cultural studies have found that gender equity in school enrollment, women's share of research jobs, and women's parliamentary representation were the most powerful predictors of cross-national variability in gender gaps in math (Else-Quest, Hyde, & Linn, 2010).

The Impact of Peers on Gender Development

As a child's social world expands outside the home, peer groups become another agency of gender development. The peer group is often thought of as the primary socializing agency of gender development and differentiation (Leaper, 1994). Peers are both the

¹ Single- and double-digit addition, subtraction, multiplication, and division.

product and the contributing producers of gender differentiation (Bandura, 1986). They instate gender differentiation by favoring same-gender playmates and making sure that their peers conform to the conduct expected of their gender.

Vigilance is needed to ensure that female teachers do not pass on their own anxieties and stereotypes to their female students (Beilock, Gunderson, Ramirez, & Levine, 2010) and that equitable classroom-level interactions are maintained (Boaler, 1997). Studies have shown that cooperative learning has been extremely effective in mathematics education for female students (Slavin, 1990; Slavin, Lake, & Groff, 2009). Still, answers from male students in a cooperative setting often prevail over those from female students, especially during dissension episodes (Wilkinson, Lindow, & Chiang, 1985).

One way that has been found to create positive experiences for female students is through personalized instructional practices. In one study, students who perceived teachers to be interested in them as individuals (personalization), who placed emphasis on investigative skills (investigation), and who felt classroom participation was important were more likely to believe in themselves as capable learners of mathematics (Forgasz & Leder, 2006). This personalization was more critical for female than for male students.

Strategic Mathematical Ability and Performance

As children mature, they proceed through a number of stages in their mathematical development. For example, they shift from physical (e.g., finger counting) to cognitive representations of numbers and operations, which decreases the load on working memory (Case et al., 1996). The acquisition of a mental "number line" and an increase in working memory capacity support a rich and abstract cognitive representation of numbers (Baroody, Tii-likainen, & Tai, 2006). Children increasingly develop meta-

cognitive abilities that enable them to know when, why, and how to use new mathematical strategies that provide flexibility in using the problem-solving and meta-strategic knowledge strategies they possess (Carr, Alexander, & Folds-Bennett, 1994; Montague & Jitendra, 2006).

Self-Regulation and Mathematics Performance

Self-regulated learners (those who take control of and evaluate their own learning and behavior) often hold incremental beliefs about intelligence (as opposed to fixed views of intelligence) and attribute their successes/failures to factors (e.g., effort expended, a particular strategy) within their control (Dweck & Leggett, 1988). Effective learners often use metacognition (thinking about one's thinking) and strategic action (planning, monitoring, and evaluating). They are highly motivated and have a high sense of self-efficacy (Butler & Winne, 1995; Perry, Phillips, & Hutchinson, 2006; Pintrich & Schunk, 2002; Winne & Perry, 2000), and they often exhibit success in and beyond school (Corno et al., 2002; Winne & Perry, 2000). Any weakness in a student's regulation of any of these areas can produce a less than optimal use of learning software. An example is "gaming the system" (Baker, Corbett, Koedinger, & Wagner, 2004), in which students perform actions to obtain the answer without seriously trying to learn the underlying concepts.

Theoretical Foundation for a Technology Environment

In this article, we describe Wayang Outpost, a multimedia-based intelligent tutoring system that is the test bed for this research (Woolf, 2009). It provides a broad range of pedagogical support while students solve mathematics problems of the type that commonly appear on standardized tests (see Figure 1; Arroyo, Beal,

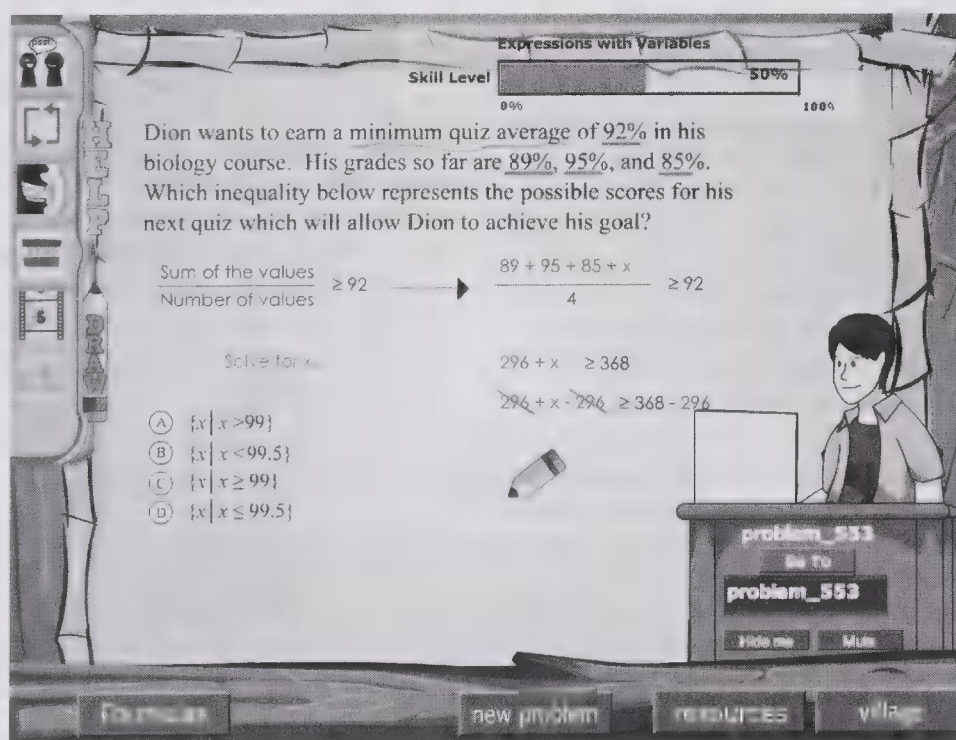


Figure 1. Wayang Outpost. In Wayang Outpost, the male learning companion uses gestures to offer advice and encouragement. Students ask for hints or to see the solution, animations, or videos.

Murray, Walles, & Woolf, 2004). Developed at the University of Massachusetts Amherst, the Wayang Tutor supports strategic and problem-solving abilities based on the theory of cognitive apprenticeship (Collins, Brown, & Newman, 1989) in which a master teaches skills to an apprentice. In this case, the expert is the computer program that assists students to learn tacit processes while the program models a solution, provides practice opportunities with the availability of scaffolded strategies based on the multimedia learning theory (Mayer, 2001), and provides metacognitive scaffolds, such as stopping to reflect on student progress.

Teaching Strategies in a Technology Environment

Wayang Outpost is particularly strong at coaching and scaffolding. It provides synchronized sound, animations, and videos that show instructors solving problems and graphic pencils to support student drawings (see Figure 1). A big part of cognitive apprenticeship is to challenge students by providing slightly more difficult problems than the learner/apprentice could accomplish by herself. Vygotsky (1978) referred to this as the zone of proximal development and suggested that fostering development within this zone led to the most rapid learning. The software provides adaptive selection of problems with increased/decreased difficulty depending on recent student success and effort (Arroyo, Cooper, Burleson, & Woolf, 2010; Corbett & Anderson, 1995).

Affective Interventions in a Technology Environment

Students' affective states and traits (e.g., frustration, boredom) can bias the outcome of any learning situation, whether human- or computer-based. Student emotions within a traditional classroom have been described as control- or value-oriented (Pekrun, 2006; Pekrun, Frenzel, Goetz, & Perry, 2007). This control-value theory is based on the premise that student appraisals of control and values are central to the arousal of achievement emotions, including activity-related emotions such as enjoyment, frustration, and boredom experienced while learning, as well as outcome emotions such as joy, hope, pride, anxiety, hopelessness, shame, and anger. Students often use coping strategies to regulate their emotions in stressful learning situations (e.g., avoidance, humor and acceptance, and negation; Eynde, de Corte, & Verschaffel, 2007).

Given the importance of affect during learning, there have been various efforts related to designing models that can automatically recognize student affect (Conati & Maclaren, 2009; D'Mello & Graesser, 2012a, 2012b; Muldner, Burleson, & VanLehn, 2010). In our prior work, a linear regression model was used to predict student emotion based on recent student behavior in the system; physical sensors (camera, seat cushion, etc.) were also used to accurately help predict students' self-reports of emotions within the software, every 5 minutes, but after a problem was complete (Arroyo, Cooper, et al., 2009; Arroyo, Woolf, Royer, & Tai, 2009; Cooper, Arroyo, & Woolf, 2011; Cooper et al., 2009). This affective model based on sensors was not used for any of the current studies in this article, which relied on students' self-reports of their emotions.

The presence of someone who cares, or at least appears to care, can make a student's experience more personal and help that student persist at a task. Brain signals imitate feelings in the body of a listener (Rapson, Hatfield, & Cacioppo, 1994); thus, a student

might register joy or sadness from someone nearby exhibiting those emotions. Empathic responses might work when students do not feel positive about the learning experience (McQuiggan, Rowe, & Lester, 2008). Thus, a computer persona that appears to enjoy math experiences could transmit positive experiences to students. Same-gender virtual characters are likely to be more effective as confidants based on research that same-gender friends are more often confidants (Reisman, 1990) and that teenagers are more intimate in same-gender friendships (Aukett, Ritchie, & Mill, 1988).

Gendered and multicultural companions in Wayang Outpost act like peers/study partners who care about a student's progress and offer support and advice (see Figure 2; Arroyo, Cooper, et al., 2009; Arroyo, Woolf, et al., 2009). Companions were designed to appear unimpressed or to simply ignore students' solutions when students did not exert effort; companions praised students who exerted effort, even if the answers were wrong. Affective characters in Wayang Outpost implemented Dweck's theory of motivation and praise (Dweck, 1999, 2002a, 2002b), which holds that students who view their intelligence as fixed and immutable (trait-based) tend to shy away from academic challenges, whereas students who believe that intelligence can be increased through effort and persistence (state-based) tend to seek out academic challenges. Praise, when delivered appropriately, can encourage students to view their intelligence as malleable and can support stable self-esteem regardless of how hard the students work. In contrast, stakeholders (e.g., teachers and parents) may lead students to accept a trait-based view of intelligence by praising intelligence, rather than effort, thus implying that success and failure depend on something beyond the students' control. Dweck's recommendations were implemented in the messages delivered by Wayang's companions. Tables 1 and 2 present a few of approximately 50 spoken messages used to motivate students and provide metacognitive help for effective problem-solving strategies. The companions speak the messages at the beginning of new problems and/or after problem-solving actions.

General Method

We investigated specific pedagogical approaches within Wayang through a series of studies conducted over the course of 10 years. Our goal was to improve both student learning and affective outcomes. The conditions in these studies varied, such as the way

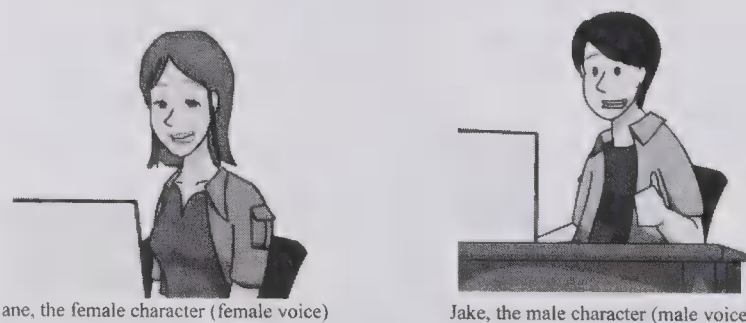


Figure 2. Gendered learning companions: Jane, the female affective learning companion, and Jake, the male affective learning companion. Characters of different racial-ethnic backgrounds were created after these studies. Jane and Jake are the two characters that were part of the studies reported in this article.

Table 1
Sample Messages Given by Learning Companions That Emphasize Effort Over Success, When Students Solve Math Problems

Answer type	Low effort	High effort
Incorrect	"We kind of rushed to answer that one. Shall we ask the computer for help? I am sure we will get it if we take the time to solve the problem."	"These are the hard questions that I like. There is an opportunity to learn. Let's click on the help button."
Correct	"That was good, however, I prefer harder questions so that we learn from the help that the computer gives, even if we get them wrong."	"Hey, congratulations! Your effort paid off, you got it right."

that help was provided or the presence of an animated character (with randomized assignment of students to conditions). However, the mathematics content remained constant, including the range of topics and other help aids (e.g., read-aloud, worked-out examples, tutorial videos).

Subjects and Procedure

Students within several mathematics classes at a variety of public schools in Massachusetts interacted with Wayang Outpost for approximately four 1-hr sessions during normal mathematics classes, during the span of 1 week. Students completed a pretest and posttest assessing their mathematics knowledge through a series of questions (approximately 15 question items per test) prior to and shortly after using Wayang, respectively. Wayang logged all student interface actions for all studies. The test varied per study depending on the knowledge units taught, which were selected depending on grade, school level, and topics the teacher chose to cover.

Instruments

Students' mathematics knowledge and problem-solving ability were assessed with instruments drawn from the Massachusetts state-based test and SAT-Math tests. Each test was composed of 15 items that were representative of knowledge units and math skills taught by Wayang Outpost. The two tests (A and B) were counterbalanced and randomly assigned, so that half the students received Test A as a pretest and half received Test B (and then reversed at posttest time); the test items were presented in random order to avoid order effects. Test items were a mix of short-answer and multiple-choice items, about half of each (see Figure 3). Two items, one easier and one harder, assessed each knowledge unit that the system was preset to cover. The items were carefully selected, so that the two tests would be fairly equivalent in diffi-

culty and would be balanced in short-answer items compared to multiple-choice items. As a result, we have not seen significant differences in performance between the two tests in any of the studies reported in this article, for students receiving either test at pretest time.

Study 1: Large-Scale Pilot (Pilot Study)

Study 1 (referred to as the pilot study) involved 139 high school students (9th–10th graders) from two schools: an urban, low-achieving school and a rural, high-achieving school. A single condition was used in which students were supposed to initiate help when they needed it. Our purpose in this study was to analyze the potential of Wayang Outpost to improve mathematics performance (no learning companions were present) and to determine behavioral predictors of mathematics learning (in terms of improvement from pretest to posttest).

Study 2: Design of Help (Help Study)

Study 2 (referred to as the help study) involved 64 students from a traditionally disadvantaged population in an urban public school (81% low income) with a large percent of Latino students. We used a between-subjects design with two experimental conditions: *tutor-initiated help* ($N = 30$) and *student-initiated help* ($N = 34$), with students randomly assigned to either condition. Students in both conditions were given access to Wayang hints that would lead them toward the solution (they clicked on a button labeled "Help"). Only students in the tutor-initiated condition were automatically offered help when they made mistakes. In the tutor-initiated condition, students were given the possibility to reject the help when offered. The main hypothesis was that tutor-initiated help would probably support students and promote higher learning more than would standard help, because it would help to regulate help-

Table 2
Sample Messages Given by Learning Companions to Train the Concept of Malleability of Intelligence

Malleability training	"Did you know that when we learn something new our brain actually changes? We form new connections inside that help us solve problems in the future. Isn't it amazing?"
Perseverance training	"Struggling in problems is actually a good thing, because it means that we are learning something new and making our minds grow."
Demythifying	"Hey, I found out that people have myths about math, like 'only some people are good at math.' The truth is that we can all be successful in math if we give it a try."

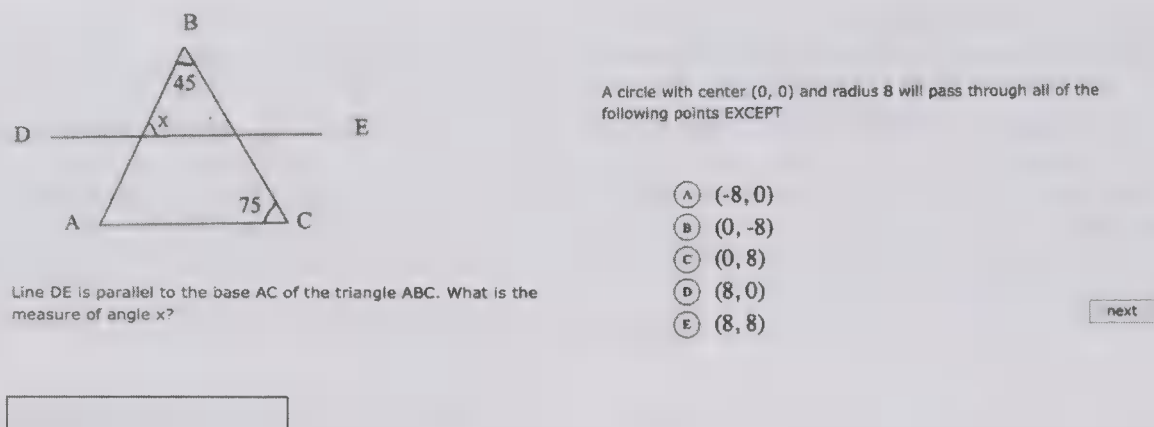


Figure 3. Two items from the math pretest and posttest instruments to assess mathematics knowledge. The item on the left is a short-answer item, to assess student knowledge of corresponding angles and internal angles of a triangle; the multiple-choice item on the right assesses student knowledge of circles and circle radius, x - y coordinates, and logical statements.

seeking behavior. As was the case with the pilot study, no learning companions were involved.

Study 3: Use of Learning Companions (LC Study)

Study 3 (referred to as the LC study) involved 233 students from a rural-area high-achieving middle school (Grades 7 and 8) and used a between-subjects design with two conditions: *Wayang with learning companions* ($N = 103$) and *standard classroom-based instruction* ($N = 100$). The hypothesis was that Wayang Outpost with learning companions would help students to improve performance outcomes more than would classroom instruction. Students were randomly assigned to the gender of the learning companion.

Study 4: Impact of Learning Companions on Student Affect (Affect Study)

Study 4 (referred to as the affect study) involved 108 students (9th–10th graders) from two rural-area high schools in Massachusetts and used a between-subjects design with two conditions: *learning companion* ($N = 72$) and *no learning companion* ($N = 36$). One goal was to see if learning companions helped to improve students' affective states while students worked with Wayang and their affective predispositions and attitudes after using Wayang, compared to those of students in the no learning companion condition. Every 5 minutes, but only after a mathematics problem was complete, the tutor asked students to self-report their cognitive-affective emotion. Students were asked to report on their interest, frustration, confidence, and excitement (e.g., "How frustrated do you feel?"). The responses were chosen from a 6-point scale. Confidence and interest were emotions with bipolar scales, with *I feel anxious* and *I am bored* at each low end, respectively, and *I feel confident* and *Very interested* at the high end.

A correlation analysis validation study over 253 students determined that those four emotions were highly similar to several activating and deactivating emotions identified by Pekrun (Arroyo, Shanabrook, Burleson, & Woolf, 2012) as measured through the Achievement Emotions Questionnaire for Mathematics (Pekrun et al., 2007). Students were randomly assigned to the conditions and to the gender of the learning companion. In addition to taking the mathematics pre- and posttests, students filled out an affective

questionnaire prior to and after using Wayang. The questionnaire assessed their affective predispositions toward mathematics learning and math problem solving, and it acted as a baseline for the four assessed emotions within the system and after the system at posttest time. The post survey also included questions about the student's perceptions of the Wayang Tutor ("Did you learn?" "Did you like it?" "Was it helpful?" "Was it friendly?").

To obtain process data on students' emotions during their interaction with Wayang, the system prompted students to report their affective state every 5 minutes, with the emotion of choice being randomly selected in the question. The affective question appeared on the screen only after a mathematics problem was complete (students were not interrupted while solving a problem) and only when the student requested a new activity.

Results

Data analyses of these four studies focused on gender differences observed before, during, and after students used Wayang. This enabled researchers to examine gender differences in subjects' affective predispositions and math performance before and after the use of the Wayang system and the affective state at each time interval, in addition to the problem-solving performance, timing, and help activity at every practice problem students saw.

Gender Differences Before Using Wayang

The differences in prior mathematics achievement for each gender before students used the tutor were small and nonsignificant across genders, across all studies, with no consistent trends in either direction for either gender. Thus, extensive data from this research suggest that before the students used the technology, no significant gender difference existed in mathematical ability.

Affective predispositions toward math problem solving in particular were assessed at pretest time through questions of the type "How [confident/frustrated] do you feel/get when solving math problems?" This gender difference became evident at the high school level for "How *confident* do you feel when solving math problems?" in the affect study in 2009 and for another study carried out in 2009; this is shown in Table 3. A gender

Table 3
Means, Standard Deviations, and Cohen's *d* Effect Sizes for Gender Differences, for Affective Predispositions Before Using the Tutoring System

Study	N girls	N boys	Confidence			Frustration			Interest			Excitement		
			Girls <i>M</i> (<i>SD</i>)	Boys <i>M</i> (<i>SD</i>)	(Girls – boys) Cohen's <i>d</i>	Girls <i>M</i> (<i>SD</i>)	Boys <i>M</i> (<i>SD</i>)	(Girls – boys) Cohen's <i>d</i>	Girls <i>M</i> (<i>SD</i>)	Boys <i>M</i> (<i>SD</i>)	(Girls – boys) Cohen's <i>d</i>	Girls <i>M</i> (<i>SD</i>)	Boys <i>M</i> (<i>SD</i>)	(Girls – boys) Cohen's <i>d</i>
Study 3, LC ^a	105	109	3.85 (1.3)	3.91 (1.3)	0.05	2.91 (1.2)	2.78 (1.3)	0.10	3.10 (1.2)	2.44 (1.3)	0.53***	2.83 (1.3)	2.25 (1.3)	0.46**
Study 4, Affect	55	52	3.07 (1.2)	3.79 (1.4)	-0.54**	3.6 (1.1)	3.09 (1.3)	0.43*	2.31 (1.0)	2.53 (1.3)	-0.19	2.4 (1.1)	2.42 (1.5)	-0.02
2010 ^b	48	44	3.17 (1.3)	3.36 (1.3)	-0.15	4.06 (1.4)	3.34 (1.4)	0.51*	2.94 (1.4)	2.86 (1.4)	0.06	2.81 (1.5)	2.57 (1.4)	-0.17
2009 ^b	70	62	3.74 (1.1)	4.18 (1.3)	-0.37*	3.44 (1.1)	2.91	0.46**	2.81	2.41	0.34	2.64	2.45	-0.16
2008 ^b	12	19	3.08 (0.9)	3.11 (1.0)	-0.03	3.42 (0.9)	2.63 (1.1)	0.77*	2.75 (1.2)	2.58 (1.2)	0.14	2.75 (1.2)	2.21 (1.1)	0.46

Note. Emotions correspond to question surveys in a 1–6 scale where 6 is the maximum value. A negative number indicates that male students had a higher outcome for the corresponding variable than female students did. Numbers in bold type indicate significant values. LC = learning companions study; Affect = affect study.

^a Indicates the only study involving middle school students instead of high school students. ^b Indicates data from a different study with high school students, which goes beyond the scope of this paper.

* Significant at $p < .05$. ** Significant at $p < .01$. *** Significant at $p < .001$.

difference was notably present for all the high school studies run by the authors since 2008 for “How *frustrated* do you get when solving math problems,” including the affect study and three previous studies involving high school students. These studies were run in different high schools in Massachusetts, with a wide variety of socioeconomic and racial-ethnic student backgrounds.

In further analyses regarding Study 4 (affect study), interactions between achievement level (median-split on math pretest score) and gender were considered. The group that reported the lowest confidence and highest frustration was low-achieving girls, who reported more negative-valenced baseline emotions than did low-achieving boys: for confidence, low-achieving girls $M = 2.71$, $SD = 1.21$; low-achieving boys $M = 3.44$, $SD = 1.34$, $F(1, 63) = 5.1$, $p = .029$. In fact, high-achieving girls reported very similar affective predispositions to those reported by low-achieving boys (e.g., for confidence, high-achieving girls $M = 3.61$, $SD = 1.08$), whereas high-achieving boys reported more positive affective predispositions than did high-achieving girls (for confidence, high-achieving boys $M = 4.35$, $SD = 1.39$).

On the other hand, the LC study involving 214 middle school students did not show a gender difference in either confidence or frustration, even though it was carried out in the same school as the affect study and the study marked as 2009 in Table 3. In fact, Table 3 shows a significant advantage for female students in both interest and excitement, which is not evident in any of the high school studies. These results suggest an important shift in gender differences in attitudes toward mathematics problem solving at the transition from middle to high school: Female students decrease their interest and excitement toward math problem solving in relation to their male peers, and they increase their frustration and decrease their confidence when solving math problems in relation to their male peers.

Similar gender differences in confidence/anxiety with high school students were found in studies carried out in Germany involving over 500 students, in which female teenagers reported more anxiety, hopelessness, and shame when learning mathematics and taking tests (Frenzel, Pekrun, & Goetz, 2007). Although it is unclear whether this affective difference at this age is because of a gender difference in teenage students' ability to report and become aware of these emotional experiences, the results in Table 3 show an important shift from middle to high school. This shift suggests that educational approaches for mathematics education must address a deteriorating affective relationship toward math problem solving, for female students in particular.

It is also important to note that all students reported alarmingly low levels of interest and excitement toward math problem solving, with both genders scoring lower than the neutral level (3.5) in all studies since year 2008. In particular, low-achieving high school students in the affect study disliked mathematics more, valued it less, had worse perception of their mathematics ability, and reported feeling worse when solving math problems than did high-achieving students (Woolf et al., 2010). We conclude that it is extremely necessary to make mathematics more interesting and exciting for all students in general and that high school girls and low-achieving students especially need support to address activating negative emotions such as frustration and anxiety.

Gender Differences While Using Wayang

While working within the rich context of tutoring systems, students display a variety of measurable behaviors (e.g., they use help, solve problems, and request hints). The combination of errors, hints, and time spent in a math practice problem enables researchers to estimate levels of mastery and engagement and to distinguish between low mastery and lack of motivation (e.g., answering quickly and incorrectly without requesting a hint is generally an indication of quick-guessing and disengagement, rather than low mastery of a skill). By combining three basic behaviors (help seeking, disengaged behavior, and affect), researchers can better understand more complex learning behaviors (Arroyo, Mehranian, & Woolf, 2010).

Help seeking and responses. Tutoring systems typically provide various forms of help and often immediately provide students with help as soon as they enter incorrect answers, in part because many students do not exhibit effective help-seeking behavior and so need scaffolding (Aleven, McLaren, Roll, & Koedinger, 2004; Aleven, Stahl, Schworm, Fischer, & Wallace, 2003). Based on prior research, Wayang Outpost encourages students to read problems, recognize if they can solve the problem, ask for help, spend time to understand the hint, and try to solve the problem. However, students do not necessarily adhere to this ideal model and tend to avoid help or abuse help (Aleven et al., 2004; Baker et al., 2004). In fact, the ways in which help seeking occurs are fairly complex; Aleven, McLaren, Roll, and Koedinger (2006) identified 57 production rules that capture both effective and ineffective help-seeking behavior. Baker et al. (2004) identified a variety of sub-optimal help seeking and other behaviors that were considered ways of gaming the system.

Adopting the gaming the system approach, we defined two specific behaviors. *Help abuse* implies that students ask for all hints quickly, so that the answer is revealed. *Quick-guessing* entails clicking fast through several attempts, either by spending only a few seconds between incorrect attempts or by making a first attempt within seconds after a problem is displayed.

One research issue was to identify how students' help-seeking behavior relates to their learning. As a part of Study 1 (pilot study), where help was available but was student initiated, we conducted a correlational analysis between students' help-seeking behavior and learning. This study did not include learning companions. Help-seeking behavior and particularly spending time on problems where help was seen (e.g., spending time thinking about and processing help and seeking deeply for more hints when necessary) were important predictors of student pre- to posttest improvement (Arroyo et al., 2004). Important gender differences were found in this study, including female students working more slowly on problems (seconds per problem) and spending more time between incorrect attempts. On average, female students invested significantly more time on hints, specifically in the total minutes on problems where help was seen (for female students, $M = 94.92$ min, $SD = 61.21$; for male students, $M = 76.28$ min, $SD = 62.05$, $p < .05$), although the genders saw similar amounts of hints per problem (for female students, $M = 0.86$ hints per problem, $SD = 0.81$; for male students, $M = 0.90$ hints per problem, $SD = .95$) and completed similar amounts of problems. This is important, because seeking help was found to be a behavior conducive to learning (Aleven et al., 2003; Arroyo et al., 2004). It

suggested that girls were using the system more productively, even though there was only a trend for girls to have higher learning in that study (pretest to posttest improvement).

In addition, girls indicated having more positive goals at posttest time and reported "seriously trying to learn" more frequently, whereas boys reported "trying to get it over with" more often. Students' motivational goals were in turn correlated with productive behaviors within the system (e.g., "seeking deeply for help" was significantly correlated with pretest to posttest improvement). We find similarities between student reports of seriously "trying to learn" with measures of learning/mastery orientation (Dweck, 1999). Girls also perceived the software as more helpful than did boys and more often reported that they would use it again. This initial study suggested that female students made more productive use of the Wayang tutoring system, taking greater advantage of its problem-solving support features than did male students and reporting more productive attitudes toward learning.

The subsequent experimental Study 2 (help study) explored the impact of having Wayang explicitly offer hints (instead of expecting students to request them), given that it was possible that many students were not making sufficient use of the available help and scaffolding. Although students in general improved more from math pretest to math posttest in the tutor-initiated help condition, gender differences were observed in students' acceptance or rejection of that help. When help was explicitly offered and tutor initiated, girls tended to accept the hints offered but boys tended to refuse them, with the final effect that female students saw more hints in total (Cohen's $d = 0.59$, $p < .05$).

In Study 4 (affect study), we analyzed the impact of learning companions for male and female students. We found that the presence of learning companions influenced productive help-seeking behaviors differentially for male and female students. Significant gender differences were observed in what we call "productive behaviors," which refers to the amount of time spent on help (Arroyo & Woolf, 2005). In this case, we compared students who received learning companions to those who did not. A significant interaction between gender and condition on "time in helped problems" suggests that, when the female learning companions were present, female students spent more time on hints (more time in problems where help was seen; Cohen's $d = 0.54$, $p < .05$). Apparently, female students were searching more deeply for help or thinking more about help than were their male counterparts.

Disengagement behavior. Tutoring systems are designed for an ideal student who behaves in a highly motivated way and tries to learn. However, metacognitive bugs do occur while students are using tutors (Aleven et al., 2004); also, students can fail to read the problem, not seek help, rush through, and in general game the system (Baker et al., 2004). Some disengaged behavior is directly related to help-seeking activities (e.g., help abuse, which involves rushing through hints until the correct answer is revealed).

In Study 1 (pilot study), girls spent more time on problems and less time quick-guessing than did boys (i.e., girls spent more time between incorrect attempts for a given problem). Moreover, boys saw more problems and girls tended to see fewer problems but spent more time on each. However, male students more frequently "abused help" by using the help to see the final answer (Cohen's $d = 0.83$, $p < .01$). Thus, boys likely rushed through the content more than did girls, suggesting they were more frequently disen-

gaged. In addition, in the same student-initiated-help condition in Study 2 (help study) that matched the condition of the pilot study, male students quick-guessed more often (Cohen's $d = 0.42$, $p < .01$).

In Study 3 (LC study), where all students were assigned to a random character, male students abused help more often than did female students (Cohen's $d = 0.59$, $p < .05$).

Last, in Study 4 (affect study), gender differences also were evidenced in disengagement behavior connected with having versus not having learning companions (Condition \times Gender interaction effects). These differences were found for actions such as help abuse, quick-guessing, and skipped problems; they indicated that girls were more engaged when companions were present. In particular, when the female learning companion was present, female students were less likely to quick-guess almost two standard deviations less than male students (Cohen's $d = 1.80$, $p < .005$). On the other hand, male students showed advantages regarding disengagement behaviors for the male character in particular: Table 4, row 10, shows that male students receiving the male character quick-guessed 1.55 standard deviations less than when receiving no character at all, thus improving their behavior.

Gender differences in affect. We observed gender differences in students' self-reported affective experiences while they used Wayang Outpost in Study 4 (affect study) and Study 3 (LC study). We carried out analyses of variance for each of the affective and behavioral dependent variables (post tutor and within tutor) shown in Table 4 for data from Study 4 (affect study). Covariates consisted of the corresponding pretest baseline variable (e.g., when analyzing confidence toward problem solving inside the tutoring system or at posttest time, we accounted for the pretest baseline confidence). Independent variables corresponded to condition (e.g., either learning companion [LC present/absent] or group [female companion/male companion/no companion]). We analyzed both main effects and interaction effects for student gender and condition over all student data.

Students in general reported more interest (less boredom) when learning companions were present than when they were absent (see

Table 4, rows 1 and 2). However, all of the results for other outcomes were gender dependent. For instance, female students significantly reported lower frustration when working with the female character, but this did not happen for male students, nor for female students who worked with the male character (see rows 3 and 4 in Table 4).

We also analyzed gender differences for these affective outcomes (note that gender differences for a single condition are not shown on Table 4). We found a gender difference for the no-learning-companion condition that indicates that male students reported more excitement than female students in the no-learning-companion condition (Cohen's $d = 0.68$, $p < .05$) and a trend for male students reporting higher levels of interest than female students in the male-character condition (Cohen's $d = 0.47$, $p < .1$). Table 4 does show a specific benefit of the Jane female character for female students on reports of excitement, compared to the control condition (see row 7 on Table 4).

These results suggest that, when the goal is to reduce students' frustration or increase excitement and interest, girls should receive the female learning companion, male students should receive the male character or no character at all, and should not receive the female character.

Gender Differences After Using Wayang

Gender differences in learning. Our studies indicate that, after use of Wayang, there are moderate effects of affective character presence improving learning gains, depending on gender and condition. Female students learned more than male students in Study 3 (LC study) when the female (Jane) character was present, and female students improved less than male students from pretest to posttest when characters were absent in the affect study.

Even when characters are absent, such as the help study and the pilot study, girls in general did not learn significantly more than boys with Wayang Outpost. But there was a trend to have higher mean learning gains and display more productive behaviors, which have been shown to be significant predictors of learning (Arroyo &

Table 4
Study 4: Effect Sizes for Posttest and Within-Tutor Emotion Self-Reports

Self-report	LC vs. control			Female character vs. control			Male character vs. control		
	All subjects	Female subjects	Male subjects	All subjects	Female subjects	Male subjects	All subjects	Female subjects	Male subjects
1. Interest within-tutor	0.15	0.03	0.10	0.16	0.06	-0.12	0.16	0.01	0.32
2. Interest post-tutor	0.29[†]	0.44	0.18	0.14	0.37	0.18	0.41	0.65[†]	0.25
3. Frustration within-tutor	-0.26	-0.68*	0.11	-0.46	-0.99***	0.00	-0.12	-0.60	0.21
4. Frustration post-tutor	-0.30	-0.48*	-0.16	-0.32	-0.57[†]	-0.16	-0.20	-0.34	-0.14
5. Confidence within-tutor	-0.04	0.06	-0.10	-0.08	0.06	-0.22	0.01	0.06	-0.02
6. Confidence post-tutor	0.13	0.26	0.04	0.16	0.41	0.04	0.10	0.09	0.14
7. Excitement within-tutor	0.11	0.58	-0.16	0.23	0.76*	-0.17	-0.01	0.34	-0.16
8. Excitement post-tutor	0.06	0.35	-0.27	-0.05	0.38	-0.27	0.14	0.51	-0.08
9. Productive behavior: Time on hints	0.18	0.36	-0.02	0.26	0.53	-0.15	0.05	-0.26	0.39
10. Disengagement: Quick-guessing	-0.02	-0.59*	-0.50	-0.10	-0.41	0.58[†]	-0.07	0.49	-1.55***

Note. This table shows effect sizes (Cohen's d) for having/not having a learning companion in general (columns 2-4), and specifically for students who got the female affective character (columns 5-7) or for students who got the male affective character (columns 8-10). In all cases, the control condition corresponds to students receiving no characters. A positive number indicates that the mean was higher for the experimental condition; a negative one indicates a higher mean for the control without affective characters. Significance level corresponds to a main effect for condition, the Fisher test of between-subjects for the corresponding analysis of variance, with pretest baseline as a covariate. Numbers in bold type indicate significant values. LC = learning companion and includes the female and male characters.

[†] Near significant at $p < .1$. * Significant at $p < .05$. *** Significant at $p < .005$.

Woolf, 2005). For instance, in the help study, all students learned more in the tutor-initiated help condition, $F(1, 63) = 6.38, p = .15$. However, female students had a tendency to improve more in that condition (female improvement from pretest to posttest, $M = 0.24$ [i.e., 24%], $SD = 0.21$; male improvement, $M = 0.12, SD = 0.24$, Cohen's $d = 0.57$). This is presumably because female students accepted more hints than did male students, thus receiving more support, and this translated into a trend for higher mean improvement.

Significantly higher learning was observed in Study 3 (LC study) by matching the gender of the companion to the gender of the student. Further analyses consisted of the influence of a gender-matched variable that would be true when the gender of student and character were the same. The result was a significant effect for matched gender at predicting learning gains, $F(1, 88) = 3.5, p = .048$, which confirms that students improve their math problem-solving performance more with a character that matches their own gender.

Further analyses of Study 2 (help study) revealed a general trend for students to learn more with the experimental version of Wayang that offered hints when incorrect attempts were made. Student learning gains revealed a main effect for condition, $F(1, 50) = 6.4, p = .015$. In addition, a significant interaction effect for Condition \times Gender \times Incoming Mathematics Ability emerged, $F(1, 50) = 5.1, p = .029$. This suggests that the control condition, which offered no help, was especially detrimental for female students of low math ability but was still beneficial for high-ability female students. No version of the help study was better for male students' learning at that point.

Gender differences in perceptions and motivation. Table 5 shows differences in student affect after using the Wayang tutor for Study 4 (affect study) in particular. We can observe significant gender differences that indicate that female students receiving the Jane character perceived the system significantly better than did those without a character (see Table 5, columns 4–6). These differences suggest that the no character condition generated better perceptions of the software for male subjects than for female subjects (column 3, row 1). It appears that female students perceived the tutor more positively when learning companions were present but that the opposite was true for male students, who clearly preferred the absence of the learning companion instead of

the presence of the female character. Interestingly, female students also perceived the system better when receiving the male character Jake than when receiving no character at all (column 8, row 1).

A general trend for the Jane character in particular suggests that students in general increased their self-concept of their ability to do mathematics and mathematics liking when receiving this character (see Table 5, column 4, rows 2 and 3). However, a more detailed analysis separating by gender (columns 5 and 6) suggests that this difference is due to a significant benefit for female students but not for male students.

Discussion and Conclusions

Although students of each gender had similar incoming mathematics ability, high school girls consistently reported lower confidence and higher frustration and anxiety toward mathematics at pretest time. On the other hand, middle school girls reported more interest and excitement toward math than did middle school boys, and they did not differ in confidence, frustration, or anxiety measurements. In contrast to these incoming factors, a variety of significant indicators suggest that female students particularly benefited from the Wayang Outpost tutor, both affectively and cognitively, and particularly when a female affective learning character was present. The results suggest a general advantage of affective learning companions for several affective outcomes for female students: reduced frustration; increased excitement; and increased perception of the software, self-efficacy in mathematics, and liking of mathematics.

In general, students reported significantly more interest (less boredom) when learning companions were used than when no learning companions were used. At the same time, students who received the female learning companion reported significantly higher self-concept and liking of mathematics at posttest time, although it seems that the difference is mostly due to a benefit for female students. Students who received the female learning companion also reported higher confidence toward problem solving in post-tutor surveys. Posttest excitement among female students was higher for those who worked with companions than for those who used no tutor; in contrast, excitement among male students was higher when companions were absent, and they quick-guessed less when characters were absent.

Table 5

Study 4 (Affect): Post-Tutor Differences in Student's Perception of the Learning Experience Between Experimental and Control Conditions

Posttest perception	LC vs. control			Female character vs. control			Male character vs. control		
	All subjects	Female subjects	Male subjects	All subjects	Female subjects	Male subjects	All subjects	Female subjects	Male subjects
1. Perception of software	0.03	0.82**	-0.65*	0.09	0.89**	-0.70*	0.23	0.78*	-0.15
2. Self-efficacy in mathematics	0.25	0.37	0.14†	0.50*	0.30†	0.16	-0.17	-0.02	-0.14
3. Liking of mathematics	0.06	0.38	-0.14	0.49†	0.34†	0.35	0.04	0.07	0.02

Note. Effect sizes (Cohen's d) are reported for the learning companion condition, either the female or the male character (columns 2-4); for the female character condition (columns 5-7); and for the male character condition (columns 8-10). Within each of these, experimental vs. control conditions are considered for all subjects first, then for female subjects alone, and then for male subjects alone. Significance level corresponds to the between-subjects Fisher test for the corresponding analysis of variance. A positive number indicates that the mean was higher for the experimental condition; a negative number indicates a higher mean for the control without affective characters. Numbers in bold type indicate significant values. LC = learning companion and includes the female and male characters.

† Marginally significant at $p < .1$. * Significant at $p < .05$. ** Significant at $p < .01$.

The Wayang tutor provided individualized and adaptive mechanisms for problem selection along with animated companions. One pedagogical approach was to guarantee student success by adjusting the difficulty of selected problems before moving on to harder problems, as specified in Arroyo, Mehranian, and Woolf (2010). As far as the studies reported here, the learning companions in particular appear to have an important impact on female students, with marginal benefits of the male character on male students. This advantage of same-sex matching can be explained in two ways: Either students developed self-efficacy via role modeling, or messages came through because of higher intimacy due to age-related same-sex friendships. Additionally, girls perceived the entire learning experience with Wayang significantly better than did boys, in particular when learning companions were present, whereas the opposite was true for boys, who reported better perception of learning when the companions were absent. Gender differences were also observed on posttests after students worked with the tutor, specifically in that girls reported higher confidence and lower frustration than did boys, in all conditions. Modeling and responding to student gender within intelligent tutors is particularly powerful, as it can improve teaching and personalize instruction at a very low cost.

We observed behavioral gender differences while students interacted with the tutor, even without digital companions, suggesting that girls accepted more help (and thus tended to learn more), had more productive behaviors conducive to learning (e.g., spent more time with help aids), and showed reduced disengagement (e.g., boys engaged in gaming more frequently, girls made fewer quick-guesses). However, girls engaged in more frequent quick-guessing when the male character was assigned to them.

Several reasons may be suggested for the less than optimal behavior of male students with the tutor. It is possible that male students avoid requesting or accepting help and hints because they are protective of their self-efficacy and they blame the tutor (external attributions) for their mistakes, thus adopting an avoidance strategy toward the software (Wigfield, 1988). This produces a suboptimal self-regulation strategy, as a more effective strategy would be to request hints on problems that they cannot solve.

Other research with educational technology supports the differential impact of technology on the two genders. For instance, Burleson and Picard (2007) found that female students experienced reduced frustration with their interactive activities involving pedagogical agents more than did male students. In an Australian study, first-year college students benefited from the use of online feedback, but male students chose not to complete the feedback session as often as did female students (Sanders et al., 2007). When the feedback session was shortened, male students' involvement increased and subjects who engaged with the feedback did improve their test scores. In related work (Gunn, French, McLeod, McSporran, & Conole, 2002), male subjects studying computer science were not as self-aware of their need for formative assessment as were their female counterparts.

This research highlights how to best support female students in intelligent learning environments, but it leaves open questions about how to support male students and the reasons for these differences. The beneficial result of characters for female students can be explained with social cognitive theory, which suggests that self-efficacy and self-regulation are related. If students feel a higher self-concept while learning, they should tend to self-

regulate better and have more productive behaviors and perceptions of the software and themselves as learners. Animated companions might have acted as role models to support self-efficacy by modeling, as suggested by Bandura and Bussey (2004). Lastly, by talking about myths in mathematics and reflecting about the meaning of making errors and the importance of perseverance, gendered companions apparently encouraged students to improve confidence in their abilities, thus raising their belief that they had what it takes to succeed.

Research such as described here may ultimately lead to nuanced recommendations about the type of individual support to provide for each gender and student's mathematics ability. Perhaps female students should work with female learning companions and male students should receive a male learning companion. Perhaps high-achieving male students should receive no learning companion at all. These recommendations cannot be made from a single experiment, but persistent results over 10 years, such as provided in this article, begin to provide a persuasive argument about the need for differential support for male and female students.

Although these results suggest some accommodations, we should be careful about making sweeping conclusions (e.g., that male students should never receive any kind of learning companions). In fact, there is evidence that low-achieving students (both male and female) benefited from affective learning companions (Woolf et al., 2010). These findings suggest that high-achieving male students did not benefit from learning companions in general and, in particular, that the female character was detrimental to several outcomes and behaviors. Further studies are needed about gender differences as students interact with advanced technology in other contexts and domains to validate the results provided here and to suggest specific strategies that work for male students in general.

References

- Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In J. C. Lester, R. M. Vicari, & F. Paraguaçu (Eds.), *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 227–239). Berlin, Germany: Springer.
- Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help-seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, 16, 101–128.
- Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning environments. *Review of Educational Research*, 73, 277–320. doi:10.3102/00346543073003277
- Arroyo, I., Beal, C., Murray, T., Waller, R., & Woolf, B. P. (2004). Web-based intelligent multimedia tutoring for high stakes achievement tests intelligent tutoring systems. In J. C. Lester, R. M. Vicari & F. Paraguaçu (Eds.), *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 142–169). Berlin, Germany: Springer.
- Arroyo, I., Cooper, D., Burleson, W., & Woolf, B. P. (2010). Bayesian networks and linear regression models of students' goals, moods, and emotions. In C. Romero, S. Ventura, M. Pechenizkiy, & R. Baker (Eds.), *Handbook of educational data mining* (pp. 323–338). Boca Raton, FL: CRC Press. doi:10.1201/b10274-26
- Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). Emotion sensors go to school. In V. Dimitrova, R. Mizoguchi, B. deBoulay, & A. C. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 17–24). Retrieved from <http://ebooks.isopress.nl/>

- Arroyo, I., Mehranian, H., & Woolf, B. P. (2010). Effort-based tutoring: An empirical approach to intelligent tutoring. In R. S. Baker, A. Merceron, & P. Pavlic (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 1–10). Retrieved from http://educationaldatamining.org/EDM2010/uploads/proc/edm2010_submission_46.pdf
- Arroyo, I., Royer, J. M., & Woolf, B. P. (2011). Using an intelligent tutor and math fluency training to improve math performance. *International Journal of Artificial Intelligence in Education*, 21, 135–152. doi:10.3222/IAI-2011-020
- Arroyo, I., Shanabrook, D. H., Burleson, W., & Woolf, B. P. (2012). Analyzing affective constructs: Emotions, attitudes and motivation. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 714–715). Berlin, Germany: Springer.
- Arroyo, I., & Woolf, B. (2005). Inferring learning and attitudes from a Bayesian network of log file data. In C. K. Looi, G. I. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education* (pp. 33–40). Retrieved from <http://ebooks.isopress.nl/>
- Arroyo, I., Woolf, B. P., Royer, J. M., & Tai, M. (2009). Affective gendered learning companions. In V. Dimitrova, R. Mizoguchi, B. de Boulay, & A. C. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 41–48). Retrieved from <http://ebooks.isopress.nl/>
- Aukett, R., Ritchie, J., & Mill, K. (1988). Gender differences in friendship patterns. *Sex Roles*, 19, 57–66.
- Baker, R., Corbett, A., Koedinger, K., & Wagner, A. (2004). Off-task behavior in the cognitive tutor classroom: When students “game the system”. In E. Dykstra-Erickson & M. Tscheligi (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 383–390). New York, NY: Association for Computing Machinery.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Bandura, A., & Bussey, K. (2004). On broadening the cognitive, motivational, and sociostructural scope of theorizing about gender development and functioning: Comment on Martin, Ruble, and Szkrybalo (2002). *Psychological Bulletin*, 130, 691–701. doi:10.1037/0033-2909.130.5.691
- Baroody, A. J., Tiilikainen, S. H., & Tai, Y. (2006). The application and development of an addition goal sketch. *Cognition and Instruction*, 24, 123–170. doi:10.1207/s1532690xci2401_3
- Beilock, S. L., Gunderson, E., Ramirez, G., & Levine, S. C. (2010). Female teachers’ math anxiety affects girls’ math achievement. *Proceedings of the National Academy of Sciences, USA*, 107, 1860–1863. doi:10.1073/pnas.0910967107
- Boaler, J. (1997). Reclaiming school mathematics: The girls fight back. *Gender and Education*, 9, 285–305. doi:10.1080/09540259721268
- Burleson, W., & Picard, R. (2007). Evidence for gender specific approaches to the development of emotionally intelligent learning companions. *IEEE Intelligent Systems Journal*, 22(4), 62–69. doi:10.1109/MIS.2007.69
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65, 245–281.
- Carr, M., Alexander, J., & Folds-Bennett, T. (1994). Metacognition and mathematics strategy use. *Applied Cognitive Psychology*, 8, 583–595. doi:10.1002/acp.2350080605
- Case, R., Okamoto, Y., Griffin, S., McKeough, A., Bleiker, C., Henderson, B., & Stephenson, K. M. (1996). The role of central conceptual structures in the development of children’s thought. *Monographs of the Society for Research in Child Development*, 61(1–2, Serial No. 246).
- Casey, M., Nuttall, R., Pezaris, E., & Benbow, C. (1995). The influence of spatial ability on gender differences in math college entrance test scores across diverse samples. *Developmental Psychology*, 31, 697–705. doi:10.1037/0012-1649.31.4.697
- Catsambis, S. (2005). The gender gap in mathematics: Merely a step function? In G. A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics* (pp. 220–245). Cambridge, England: Cambridge University Press.
- Cherry, L. (1975). The preschool teacher–child dyad: Sex differences in verbal interaction. *Child Development*, 46, 532–535.
- Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Erlbaum.
- Conati, C., & Maclaren, H. (2009). Modeling user affect from causes and effects. In G.-J. Houben, G. I. McCalla, F. Pianesi, & M. Zancanaro (Eds.), *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization* (pp. 4–15). Berlin, Germany: Springer.
- Cooper, D. G., Arroyo, I., & Woolf, B. (2011). Actionable affective processing for automatic tutor interventions. In R. Calvo & S. D’Mello (Eds.), *New perspectives on affect and learning technologies* (pp. 127–140). New York, NY: Springer. doi:10.1007/978-1-4419-9625-1_10
- Cooper, D., Arroyo, I., Woolf, B., Muldner, K., Burleson, W., & Christopherson, R. (2009). Sensors model student self concept in the classroom. In G. J. Houben, G. I. McCalla, F. Pianesi, & M. Zancanaro (Eds.), *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization* (pp. 30–41). Berlin, Germany: Springer.
- Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4, 253–278. doi:10.1007/BF01099821
- Corno, L., Cronbach, L. J., Kupermintz, H. K., Lohman, D. H., Mandinach, E. B., Porteus, A., & Talbert, J. E. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Mahwah, NJ: Erlbaum.
- Dehaene, S., Spelke, L., Pinel, P., Stanescu, R., & Tsivkin, S. (1999, May 7). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, 284, 970–974. doi:10.1126/science.284.5416.970
- D’Mello, S., & Graesser, A. (2012a). Dynamics of affective states during complex learning. *Learning and Instruction*, 22, 145–157. doi:10.1016/j.learninstruc.2011.10.001
- D’Mello, S., & Graesser, A. (2012b). Language and discourse are powerful signals of student emotions during tutoring. *IEEE Transactions on Learning Technologies*, 5, 304–317. doi:ieeecomputersociety.org/10.1109/TLT.2012.10
- Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia, PA: Psychology Press.
- Dweck, C. S. (2002a). Beliefs that make smart people dumb. In R. J. Sternberg (Ed.), *Why smart people do stupid things* (pp. 24–41). New Haven, CT: Yale University Press.
- Dweck, C. S. (2002b). Messages that motivate: How praise molds students’ beliefs, motivation, and performance (in surprising ways). In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 37–60). San Diego, CA: Academic Press. doi:10.1016/B978-012064455-1/50006-3
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256–273. doi:10.1037/0033-295X.95.2.256
- Ebbeck, M. (1984). Equity for boys and girls: Some important issues. *Early Child Development and Care*, 18, 119–131. doi:10.1080/0300443840180106
- Eccles, J. S. (1987). Gender roles and women’s achievement-related decisions. *Psychology of Women Quarterly*, 11, 135–172. doi:10.1111/j.1471-6402.1987.tb00781.x

- Eccles, J. S. (1989). Bringing young women to math and science. In M. Crawford & M. Gentry (Eds.), *Gender and thought* (pp. 36–58). New York, NY: Springer-Verlag. doi:10.1007/978-1-4612-3588-0_3
- Eccles, J., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self and task perceptions during elementary school. *Child Development*, 64, 830–847. doi:10.2307/1131221
- Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics and gender equity: A meta-analysis. *Psychological Bulletin*, 136, 103–127. doi:10.1037/a0018053
- Eynde, P., de Corte, E., & Verschaffel, L. (2007). Student emotions: A key component of self-regulated learning? In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 179–198). San Diego, CA: Academic Press.
- Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, 18, 850–855. doi:10.1111/j.1467-9280.2007.01990.x
- Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher*, 27, 6–11.
- Fitzgerald, L. F., & Crites, J. O. (1980). Toward a career psychology of women: What do we know? What do we need to know? *Journal of Counseling Psychology*, 27, 44–62. doi:10.1037/0022-0167.27.1.44
- Forgasz, H. J., & Leder, G. C. (2006). Academic life: Monitoring work patterns and daily activities. *Australian Educational Researcher*, 33, 1–22. doi:10.1007/BF03246278
- Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Girls and mathematics—A “hopeless” issue? A control-value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education*, 22, 497–514. doi:10.1007/BF03173468
- Gunn, C., French, S., McLeod, H., McSporry, M., & Conole, G. (2002). Gender issues in computer-supported learning. *Research in Learning Technology*, 10, 32–44. doi:10.1080/0968776020100106
- Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008, July 25). Gender similarities characterize math performance. *Science*, 321, 494–495. doi:10.1126/science.1160364
- Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, 104, 53–69. doi:10.1037/0033-2909.104.1.53
- Leaper, C. (1994). *Childhood gender segregation: Causes and consequences*. San Francisco, CA: Jossey-Bass.
- Mayer, R. E. (2001). *Multimedia learning*. New York, NY: Cambridge University Press. doi:10.1017/CBO9781139164603
- McQuiggan, S., Rowe, J., & Lester, J. (2008). The effects of empathetic virtual characters on presence in narrative-centered learning environments. In M. Czerwinski, A. M. Lund, & D. S. Tan (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1511–1520). New York, NY: Association for Computing Machinery.
- Montague, M., & Jitendra, A. K. (2006). *Teaching mathematics to middle school students with learning difficulties*. New York, NY: Guilford Press.
- Muldner, K., Burleson, W., & VanLehn, K. (2010). “Yes!”: Using tutor and sensor data to predict moments of delight during instructional activities. In P. De Bra, A. Kobsa, & D. Chin (Eds.), *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization* (pp. 159–170). Berlin, Germany: Springer.
- National Council of Teachers of Mathematics. (2008). Equity in mathematics education—A position of the National Council of Teachers of Mathematics. Retrieved from <http://www.nctm.org/about/content.aspx?id=13490>
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18, 315–341. doi:10.1007/s10648-006-9029-9
- Pekrun, R., Frenzel, A., Goetz, T., & Perry, R. (2007). The control-value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schultz & R. Pekrun (Eds.), *Emotion in education* (pp. 13–36). San Diego, CA: Academic Press. doi:10.1016/B978-012372545-5/50003-4
- Perry, N. E., Phillips, L., & Hutchinson, L. R. (2006). Preparing student teachers to support for self-regulated learning. *Elementary School Journal*, 106, 237–254. doi:10.1086/501485
- Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications*. Englewood Cliffs, NJ: Prentice Hall Merrill.
- Rapson, R. L., Hatfield, E., & Cacioppo, J. T. (1994). *Emotional contagion*. New York, NY: Cambridge University Press.
- Reisman, J. M. (1990). Intimacy in same-sex friendships. *Sex Roles*, 23, 65–82. doi:10.1007/BF00289880
- Royer, J. M., Tronsky, L. N., Chan, Y., Jackson, S. J., & Marchant, H. G. (1999). Math fact retrieval as the cognitive mechanism underlying gender differences in math test performance. *Contemporary Educational Psychology*, 24, 181–266. doi:10.1006/ceps.1999.1004
- Royer, J. M., & Walles, R. (2007). Influences of gender, motivation and socioeconomic status on mathematics performance. In D. B. Berch & M. Mazzocco (Eds.), *Why is math so hard for some children* (pp. 349–368). Baltimore, MD: Brookes.
- Sanders, K., Hill, J., Meyer, J., Fyfe, G., Fyfe, S., Ziman, M., & Koehler, N. (2007). Gender and engagement in automated online test feedback in first year human biology. In R. J. Atkinson, C. McBeath, S. Soong, & C. Cheers (Eds.), *Proceedings ascilite Singapore 2007* (pp. 909–912). Retrieved from <http://www.ascilite.org.au/conferences/singapore07/procs/sanders-poster.pdf>
- Sevo, R., & Chubin, D. E. (2010). *Lessons-learned from “Extension Services” grantees 2005–2009 “Extension Services” Grantees (NSF Research on Gender in Science and Engineering Program): A national view*. Washington, DC: American Association for the Advancement of Science and Center for Advancing Science and Engineering Capacity.
- Slavin, R. E. (1990). *Cooperative learning: Theory, research, and practice*. Englewood Cliffs, NJ: Prentice Hall.
- Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, 79, 839–911. doi:10.3102/0034654308330968
- Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, 35, 4–28. doi:10.1006/jesp.1998.1373
- Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wigfield, A. (1988). Children's attributions for success and failure: Effects of age and attentional focus. *Journal of Educational Psychology*, 80, 76–81. doi:10.1037/0022-0663.80.1.76
- Wilkinson, L. C., Lindow, J., & Chiang, C. P. (1985). Sex differences and sex segregation in students' small-group communication. In L. C. Wilkinson & C. B. Marrett (Eds.), *Gender influences in classroom interaction* (pp. 185–207). New York, NY: Academic Press.
- Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. B. Pintrich & M. Seidner (Eds.), *Handbook of self-regulation* (pp. 531–566). Orlando, FL: Academic Press. doi:10.1016/B978-012109890-2/50045-7
- Woolf, B. (2009). *Building intelligent interactive tutors: Bridging theory and practice*. San Francisco, CA: Elsevier.
- Woolf, B. P., Arroyo, I., Muldner, K., Burleson, W., Cooper, D., Dolan, R., & Christopherson, R. M. (2010). The effect of motivational learning companions on low achieving students and students with disabilities. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 327–337). Berlin, Germany: Springer.

Received December 16, 2011

Revision received February 12, 2013

Accepted February 17, 2013 ■

A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on K–12 Students' Mathematical Learning

Saiying Steenbergen-Hu and Harris Cooper
Duke University

In this study, we meta-analyzed empirical research of the effectiveness of intelligent tutoring systems (ITS) on K–12 students' mathematical learning. A total of 26 reports containing 34 independent samples met study inclusion criteria. The reports appeared between 1997 and 2010. The majority of included studies compared the effectiveness of ITS with that of regular classroom instruction. A few studies compared ITS with human tutoring or homework practices. Among the major findings are (a) overall, ITS had no negative and perhaps a small positive effect on K–12 students' mathematical learning, as indicated by the average effect sizes ranging from $g = 0.01$ to $g = 0.09$, and (b) on the basis of the few studies that compared ITS with homework or human tutoring, the effectiveness of ITS appeared to be small to modest. Moderator analyses revealed 2 findings of practical importance. First, the effects of ITS appeared to be greater when the interventions lasted for less than a school year than when they lasted for 1 school year or longer. Second, the effectiveness of ITS for helping students drawn from the general population was greater than for helping low achievers. This finding draws attentions to the issue of whether computerized learning might contribute to the achievement gap between students with different achievement levels and aptitudes.

Keywords: intelligent tutoring systems, effectiveness, mathematical learning, meta-analysis, achievement

Intelligent tutoring systems (ITS) are computer-assisted learning environments created using computational models developed in the learning sciences, cognitive sciences, mathematics, computational linguistics, artificial intelligence, and other relevant fields. ITS often are self-paced, learner-led, highly adaptive, and interactive learning environments operated through computers. ITS are adaptive in that they adjust and respond to learners with tasks or steps to suit learners' individual characteristics, needs, or pace of learning (Shute & Zapata-Rivera, 2007).

ITS have been developed for mathematically grounded academic subjects, such as basic mathematics, algebra, geometry, and statistics (Cognitive Tutor: Anderson, Corbett, Koedinger, & Pelletier, 1995; Koedinger, Anderson, Hadley, & Mark, 1997; Ritter, Kulikowich, Lei, McGuire, & Morgan, 2007; AnimalWatch: Beal, Arroyo, Cohen, & Woolf, 2010; ALEKS: Doignon & Falmagne, 1999); physics (Andes, Atlas, and Why/Atlas: VanLehn et al., 2002, 2007); and computer science (dialogue-based intelligent tutoring systems: Lane & VanLehn, 2005; ACT Programming Tutor: Corbett, 2001). Some ITS assist with the learning of reading (READ 180: Haslam, White, & Klinge, 2006; iSTART: McNamara, Levinstein, & Boonthum, 2004), writing (R-WISE writing tutor: Rowley, Carlson, & Miller, 1998), economics (Smithtown: Shute & Glaser, 1990), and research methods (Research Methods

Tutor: Arnott, Hastings, & Allbritton, 2008). There are also ITS for specific skills, such as metacognitive skills (see Aleven, McLaren, & Koedinger, 2006; Conati & VanLehn, 2000). The use of ITS as an educational tool has increased considerably in recent years in U.S. schools. Cognitive Tutor by Carnegie Learning, for example, was used in over 2,600 schools in the United States as of 2010 (What Works Clearinghouse, 2010a).

ITS are developed so as to follow the practices of human tutors (Graesser, Conley, & Olney, 2011; Woolf, 2009). They are expected to help students of a range of abilities, interests, and backgrounds. Research suggests that expert human tutors can help students achieve learning gains as large as two sigmas (Bloom, 1984). Although not as effective as what Bloom (1984) found, a recent meta-review by VanLehn (2011) found that human tutoring had a positive impact of $d = 0.79$ on students' learning.

ITS track students' subject domain knowledge, learning skills, learning strategies, emotions, or motivation in a process called *student modeling* at a level of fine-grained detail that human tutors cannot (Graesser et al., 2011). ITS can also be distinguished from computer-based training, computer-assisted instruction (CAI), and e-learning. Specifically, given their enhanced adaptability and power of computerized learning environments, ITS are considered superior to computer-based training and CAI in that ITS allow an infinite number of possible interactions between the systems and the learners (Graesser et al., 2011). VanLehn (2006) described ITS as tutoring systems that have both an outer loop and an inner loop. The outer loop selects learning tasks; it may do so in an adaptive manner (i.e., select different problem sequences for different students), on the basis of the system's assessment of each individual student's strengths and weaknesses with respect to the targeted learning objectives. The inner loop elicits steps within each task (e.g., problem-solving steps) and provides guidance with respect to

This article was published Online First September 9, 2013.

Saiying Steenbergen-Hu and Harris Cooper, Department of Psychology & Neuroscience, Duke University.

Correspondence concerning this article should be addressed to Saiying Steenbergen-Hu, Department of Psychology & Neuroscience, Duke University, 417 Chapel Drive, Box 90086, Durham, NC 27708-0086. E-mail: ss346@duke.edu

these steps, typically in the form of feedback, hints, and error messages. In this regard, as VanLehn (2006) noted, ITS are different from CAI, computer-based training, or web-based homework in that the latter lack of an inner loop. ITS are one type of e-learning that can be self-paced or instructor directed, encompassing all forms of teaching and learning that are electronically supported, through the Internet or not, in the form of texts, images, animations, audios, or videos.

The growth of ITS and the accumulation of evaluation research justify a meta-analysis of the effectiveness of ITS on students' mathematical learning for the following three reasons. First, several reviews of the impact of ITS on reading already exist (Becker, 1992; Blok, Oostedam, Otter, & Overmaat, 2002; Kulik, 2003). Most recently, Cheung and Slavin (2012) reviewed the effects of educational technology on K–12 students' reading achievement, relative to traditional instructional methods. They found an average standardized mean difference of 0.16 favoring the educational technology. No similar review regarding ITS with a focus on math has been carried out.

Second, much research on the effectiveness of math ITS has accumulated over the last two decades. Without rigorous summarization, this literature appears confusing in its findings. For example, Koedinger et al. (1997) found that students tutored by Cognitive Tutor showed extremely high learning gains in algebra compared with students who learned algebra through regular classroom instruction. Shneyderman (2001) found that, on average, students who learned algebra through Cognitive Tutor scored 0.22 standard deviations above their peers who learned algebra in traditional classrooms but only scored 0.02 standard deviations better than their comparison peers on the statewide standardized test. However, Campuzano, Dynarski, Agodini, and Rall (2009) found that sixth grade students who were taught math with regular classroom instruction throughout a school year outperformed those who were in regular class 60% of the school year and spent the other 40% of class time learning math with ITS, indicated by an effect size of -0.15 . Thus, there is a need to gather, summarize, and integrate the empirical research on math ITS, to quantify their effectiveness and to search for influences on their impact.

Third, there has been increased attention in recent years on the effectiveness of math ITS for students' learning. The What Works Clearinghouse (WWC) has completed several evidence reviews on some math ITS products. For example, the WWC produced four reviews on Carnegie Learning's Cognitive Tutor (i.e., the WWC, 2004, 2007, 2009, 2010a). The WWC also reviewed the evidence on Plato Achieve Now (see WWC, 2010b). The WWC reviews, however, did not include all math ITS. Our literature search identified more than a dozen intelligent tutoring system products designed to help students' mathematical learning. And, important for our effort, the WWC reviews did not examine factors that might influence the direction and magnitude of the ITS effect. In contrast, our effort does not focus on specific intelligent tutoring system programs but on their general effectiveness and on the factors that moderate their impact.

In sum then, a number of questions regarding the effectiveness of ITS can be addressed by a meta-analysis. Most broadly, it can estimate the overall average effectiveness of ITS relative to other types of instruction on students' mathematical learning. But more specific questions also may be answerable. For example, a meta-analysis can explore what kind of settings ITS work best in, as well

as for what types of student populations. By using information across as well as within primary studies, a meta-analysis provides a useful quantitative strategy for answering these questions.

Method

Study Inclusion and Exclusion Criteria

For studies to be included in this meta-analysis, the following eight criteria had to be met:

1. Studies had to be empirical investigations of the effects of ITS on learning of mathematical subjects. Secondary data analyses and literature reviews were excluded.
2. Studies had to be published or reported during the period from January 1, 1990, to June 30, 2011, and had to be available in English.
3. Studies had to focus on students in grades K–12, including high achievers, low achievers, and remedial students. However, studies focusing exclusively on students with learning disabilities or social or emotional disorders (e.g., students with attention-deficit/hyperactivity disorder) were excluded.
4. Studies had to measure the effectiveness of ITS on at least one learning outcome. Common measurements included standardized test scores, modified standardized test scores, course grades, or scores on tests developed by researchers.
5. Studies had to have used an independent comparison group. Comparison conditions included regular classroom instruction, human tutoring, or homework. Studies without a comparison group or those with one-group pretest–posttest designs were excluded.
6. Studies had to use randomized experimental or quasi-experimental designs. If a quasi-experimental design was used, evidence had to be provided that the treatment and comparison groups were equivalent at baseline (see WWC, 2008). Studies with a significant difference between the treatment and comparison groups prior to the ITS intervention were excluded, unless information was available for us to calculate effect sizes that would take into account the prior difference.
7. Studies had to have at least eight subjects in treatment and comparison groups, respectively. Studies with sample sizes less than eight in either group were excluded.
8. Studies had to provide the necessary quantitative information for the calculation or estimation of effect sizes.

Study Search

We used the following procedures to locate studies: (a) a search of abstracts in electronic databases including ERIC, PsycINFO, Proquest Dissertation and Theses, Academic Search Premier,

Econlit With Full Text, PsycARTICLES, SocINDEX With Full Text, and Science Reference Center; (b) Web searches using the Google and Google Scholar search engines; (c) a manual examination of reference and bibliography lists of the relevant studies; and (d) personal communications with 18 ITS research experts who had been the first author on two or more ITS studies during the past 20 years.

We used a wide variety of search terms to ensure our searches would identify as many relevant studies as possible. Although some researchers have used the term *intelligent tutoring systems*, many others have used a wide variety of alternative terms, for example, *computer-assisted tutoring*, *computer-based tutoring*, *artificial tutoring*, or *intelligent learning environments*. Therefore, we also used the terms *intelligent tutor**, *artificial tutor**, *computer tutor**, *computer-assisted tutor**, *computer-based tutor**, *intelligent learning environment**, *computer coach**, *online-tutor**, *keyboard tutor**, *e-tutor**, *electronical tutor**, and *web-based tutor**. After concluding these searches, we began to focus on math ITS.

We found that some math ITS studies could not be retrieved through the search keywords above and some ITS studies are locatable only through the use of particular ITS names. The reference list of Graesser et al.'s (2011) introduction to ITS, for instance, indicates that large numbers of studies are exclusively connected with particular ITS programs, such as Cognitive Tutor, AutoTutor, or CATO. Dynarski et al. (2007) evaluated the effectiveness of three mathematical educational software programs for sixth graders (i.e., Larson Pre-Algebra, Achieve Now, and iLearn Math) and three software programs for ninth graders (i.e., Cognitive Tutor Algebra, Plato Algebra, and Larson Algebra). All of these educational software programs were actually ITS products. However, we found our previous search only caught studies of Cognitive Tutor, the most widely used and studied ITS, and we missed all studies of the other software. This was also the case for the educational software evaluated by Campuzano et al. (2009). Therefore, we used the names of some major software programs reported in Graesser et al. (2011), Dynarski et al. (2007), and Campuzano et al. (2009) and conducted a third search in ERIC and PsycINFO. No new qualified studies were found. However, by screening all of the studies in the WWC reviews of Cognitive Tutor and Plato Achieve Now (i.e., WWC, 2004, 2007, 2009, 2010a, 2010b), we found five additional studies that qualified for inclusion. In summary, our search concluded with 26 qualified reports evaluating the effectiveness of ITS on K–12 students' mathematical learning.

Study Coding

We designed a detailed coding protocol to guide the study coding and information retrieval. The coding protocol covered studies' major characteristics, which included (a) the basic features of the study reports (e.g., whether the study was published or unpublished and when it was conducted), (b) research design features (e.g., sample sizes; whether the study used a randomized or quasi-experimental design), (c) the contexts of intervention (e.g., subject matter; whether the study compared ITS with regular classroom instruction, human tutors, or other education interventions; the duration of ITS intervention), and (d) the study outcomes (e.g., what and how outcomes were measured; when the assessments took place; the magnitudes and direction of the effect sizes).

Two coders independently coded the major features of each study, except the study outcomes, and then met together to check the accuracy of the coding. If there was a disagreement in coding, the two coders discussed and reexamined the studies to settle on the most appropriate coding. If the disagreement could not be resolved, the second author was consulted. The first author coded the study outcomes and then discussed the codes with the second author. The major specific variables coded are described later along with the study results.

Effect Size Calculation

We used Hedges' g , a standardized mean difference between two groups, as the effect size index for this meta-analysis. The preference for Hedges' g over other standardized-difference indices, such as Cohen's d and Glass's Δ , is due to the fact that Hedges' g can be corrected to reduce the bias that may arise when the sample size is small (i.e., $n < 40$; Glass, McGaw, & Smith, 1981). Hedges' g was chosen for this meta-analysis because the samples in many ITS studies are small.

Hedges' g was calculated by subtracting the mean of the comparison condition from that of the ITS tutoring condition and dividing the difference by the average of the two groups' standard deviations. A positive g indicates that students tutored by ITS achieved more learning gains than did those in the comparison condition. In cases for which only inference test results were reported but no means and standard deviations were available, g was estimated from the inferential statistics, such as t , F , or p values (Wilson & Lipsey, 2001). For studies that did not report specific values of inferential statistics, we assumed a conservative value for effect size calculation. For example, if a study reported a statistically significant difference between the ITS and the comparison condition with $p < .01$, we assumed a p value of .01 for effect size calculation.¹

We calculated unadjusted effect sizes for a study if it only reported the ITS and comparison groups' mean posttest scores, standard deviation, and sample sizes. Unadjusted effect sizes did not take into account other variables that might have had an impact on the outcomes. For some studies, in addition to unadjusted effect sizes, adjusted effect sizes were also extracted. We called them *adjusted effect sizes* because they were calculated after adjusting or controlling for other variables, such as pretest scores. In some cases, adjusted effect sizes were based on means and standard deviations of gain scores (i.e., posttests – pretests), whereas in other cases they were based on covariance-adjusted means and standard deviations. For studies that reported descriptive statistics of both pretests and posttests, as suggested by D. B. Wilson (personal communication, April 18, 2011), adjusted effect sizes were the differences between posttest and pretest effect sizes and their variances were the sum of posttest and pretest effect sizes variances.

¹ This was the case for only one study (i.e., Shneyderman, 2001), in which the effect size for one of the three outcomes was calculated by assumed $p = .01$ when the study reported $p < .01$. Because the effect size representing this study was an average of all three effect sizes from three outcomes, there was a minimal possibility that this would lead to an underestimation of the overall effect sizes in this meta-analysis.

Independent Studies, Samples, and Effect Sizes

To address effect size dependency issues, we used independent samples as the unit of analysis. Each independent sample is not the equivalent of a separate research report. One report could contain two or more independent studies. For example, we coded Beal et al. (2010) as two independent studies, each based on a different sample. The 26 reports contained 34 independent studies based on 34 independent samples. Table 1 presents the major features of all 31 studies in which ITS were compared with regular classroom instruction.

We used a shifting unit of analysis approach (Cooper, 2010) to further address possible dependencies among effect sizes. The benefits of the shifting unit of analysis approach are that it allows us to retain as much data as possible while ensuring a reasonable degree of independence among effect sizes. With this approach, effect sizes were first extracted for each outcome as if they were independent. For example, if a study with one independent sample used both a standardized test and a course grade to measure students' learning, two separate effect sizes were calculated. When estimating the overall average effect of ITS, these two effect sizes were averaged so that the sample contributed only one effect size to the analysis. However, when conducting a moderator analysis to investigate whether the effects of ITS vary as a function of the type of outcome measures, this sample contributed one effect size to the category of standardized test and one to that of course grade.

Data Analysis

We used the Comprehensive Meta-Analysis (Borenstein, Hedges, Higgins, & Rothstein, 2006) software for data analysis. Before the analyses, we conducted Grubbs (1950) tests to examine whether there were statistical outliers among the effect sizes and sample sizes. We conducted the meta-analysis using a weighting procedure and with both fixed-effect and random-effects models (Cooper, 2010). A fixed-effect model functions with the assumption that there is one true effect in all of the studies included in a meta-analysis and the average effect size will be an estimate of that value. A fixed-effect model is suited to drawing conditional inferences about the observed studies. However, it is less well suited to making generalizations to the population of studies from which the observed studies are a sample (Konstantopoulos & Hedges, 2009). A random-effects model assumes that there is more than one true effect and the effect sizes included in a meta-analysis are drawn from a population of effects that can differ from each other.

Two approaches were used to assess publication bias. First, a funnel plot was visually inspected. The suggestion of missing studies on the left side of the distribution indicated the possible presence of publication bias. Duval and Tweedie's (2000) trim-and-fill procedure (available as part of the Comprehensive Meta-Analysis software) was then used to further assess and adjust for publication bias. Through this procedure, unmatched observations were removed (trimmed) from the data distribution and additional values were imputed (filled) for projected missing studies. Then, average effect sizes are recalculated.

Moderator Analyses

Testing for moderators was conducted on the groups of effect sizes that had a high degree of heterogeneity (Cooper, Hedges, &

Valentine, 2009). The purpose of testing for moderators was to identify variables associated with certain features of the primary studies that might be significantly associated with the effectiveness of ITS.

Results

The literature search located 26 reports that met our study inclusion criteria. The reports appeared between 1997 and 2010. The sample sizes in the reported studies ranged from 18 to 17,164. The 26 reports provided 65 effect sizes. Forty-seven effect sizes were unadjusted, meaning they were calculated from posttest outcome measures and did not control for variables other than the ITS treatment, which might have influenced the outcome measures; 18 were adjusted effect sizes, which were calculated after controlling for other confounding variables, such as pretest scores.

As mentioned in the Method section, to address effect size dependency issues, we used independent studies (i.e., samples) as the unit of analysis. The 26 reports contained 34 independent studies based on 34 independent samples. Of the 34 independent studies, 31 compared ITS with regular classroom instruction. These 31 studies provided 61 effect sizes (see Table 1). In general, these 31 independent studies compared learning outcomes of instructions with an ITS component to those without one. Specifically, this comparison refers to four types of comparison situations. First, a large portion of the studies, for example, studies of Cognitive Tutor compared the learning gains of students who learned through instruction in which Cognitive Tutor was a significant part of regular classroom instruction with the learning gains of students who learned through traditional classroom instruction in which no Cognitive Tutor was involved. In such studies, interventions usually lasted for one school year or one semester during which students in the experimental groups generally spent 60% of their time in regular classroom learning and 40% of their time in the computer lab using Cognitive Tutor; students in the control groups spent 100% of their time in regular classrooms. Second, some studies compared students who learned solely through using ITS with those who learned in regular classroom instruction (e.g., Arroyo, Woolf, Royer, Tai, & English, 2010; Beal et al., 2010, Study 2; Wallis, 2005). Interventions in these studies usually lasted for just a few days. Third, two studies compared students' learning in conditions in which ITS partially took teacher's responsibilities (e.g., giving students guidance or feedback) to students' learning in conditions in which they received guidance or feedback from teachers (i.e., Hwang, Tseng, & Hwang, 2008; Stankov, Rosic, Zitko, & Grubisic, 2008). Interventions in these studies lasted for several weeks to one semester. Last, one study compared students who used ITS as a supplement in addition to regular classroom instruction with students who learned through regular classroom instruction without using ITS as a supplement (i.e., Biesinger & Crippen, 2008). Intervention in this study lasted for one semester. Because the comparison conditions in all four types of situations above involved either regular classroom instruction or teachers' efforts, we grouped them together as ITS being compared with regular classroom instruction.

Two independent studies (i.e., Mendicino, Razzaq, & Hefferman, 2009; Radwan, 1997) provided information on the effects of ITS on mathematical learning relative to that of homework assignments. One independent study (i.e., Beal et al., 2010) compared

Table 1
Studies Comparing Intelligent Tutoring Systems With Regular Classroom Instruction

Study (independent sample)	ITS name	Subject	ITS duration	Sample size ^a	Sample achievement level	Schooling level	Research design	Year of data collection	Counter- balanced testing	Report type	Unadjusted ES ^b	Adjusted ES ^c
Arbuckle (2005)	Cognitive Tutor Algebra I	Algebra	Short term	111	General	High	Quasi-experimental	ng	No	Nonjournal	0.78	0.67
Arroyo et al. (2010)	Math Facts Retrieval Training and Wayang Outpost AnimalWatch	Basic math	Short term	250	General	Middle	Quasi-experimental	2006–2010	Yes	Journal	0.39	
Beal et al. (2010) (2)	Wayang Outpost	Basic math	Short term	202	General	Middle	Quasi-experimental	ng	ng	Journal	–0.26	
Beal et al. (2007)	Wayang Outpost	Basic math	Short term	28	General	High	Quasi-experimental	ng	Yes	Journal		0.55
Biesinger & Crippen (2008) (1)	Online Remediation software	Basic math	Semester	3,566	Low achievers	High	Quasi-experimental	2003–2005	No	Journal	0.22	0.13
Biesinger & Crippen (2008) (2)	Online Remediation software	Basic math	Semester	17,164	General	High	Quasi-experimental	2003–2005	No	Journal	0.16	
Cabalo & Vu (2007)	Cognitive Tutor Algebra I	Algebra	One semester	364	General	Ng	True experimental	2006–2010	No	Nonjournal	–0.22	0.03
Cabalo, Ma, & Jaciw (2007)	Cognitive Tutor Bridge to Algebra Curriculum	Algebra	One school year	576	General	Ng	True experimental	2006–2010	No	Nonjournal	0.09	0.05
Campuzano et al. (2009) (1)	Larson Pre-Algebra and Achieve Now	Basic math	One school year	659	General	Middle	True experimental	2006–2010	Yes	Nonjournal	–0.23	–0.12
Campuzano et al. (2009) (2)	Cognitive Tutor Algebra I and Larson Algebra I	Algebra	One school year	534	General	High	True experimental	2006–2010	Yes	Nonjournal	0.09	0.09
Carnegie Learning (2001a)	Cognitive Tutor Algebra I	Algebra	One school year	293	General	High	True experimental	Before 2003	No	Nonjournal	0.45	
Carnegie Learning (2001b) (1)	Cognitive Tutor Math	Basic math	One school year	132	General	Middle	Quasi-experimental	Before 2003	No	Nonjournal		0.00
Carnegie Learning (2001b) (2)	Cognitive Tutor Math	Basic math	One school year	174	General	Middle	Quasi-experimental	Before 2003	No	Nonjournal		–0.23
Carnegie Learning (2002) (1)	Cognitive Tutor	Algebra	One school year	58	General	Middle	Quasi-experimental	Before 2003	No	Nonjournal	–0.66	
Carnegie Learning (2002) (2)	Cognitive Tutor	Algebra	One school year	80	General	High	Quasi-experimental	Before 2003	No	Nonjournal	–0.24	
Dynarski et al. (2007) (1)	Larson Pre-Algebra, Achieve Now, and iLearn Math	Basic math	One school year	3,136	General	Middle	True experimental	2003–2005	Yes	Nonjournal	0.05	0.14
Dynarski et al. (2007) (2)	Cognitive Tutor Algebra, Plato Algebra, and Larson Algebra	Algebra	One school year	1,404	General	High	True experimental	2003–2005	Yes	Nonjournal	–0.23	–0.07

Table 1 (continued)

Study (independent sample)	ITS name	Subject	ITS duration	Sample size ^a	Sample achievement level	Schooling level	Research design	Year of data collection	Counter- balanced testing	Report type	Unadjusted ES ^b	Adjusted ES ^c
Hwang, Tseng, & Hwang (2008)	Intelligent Tutoring, Evaluation and Diagnosis	Basic math	Semester	76	General	ng	True experimental	ng	No	Journal	0.75	
Koedinger (2002)	Cognitive Tutor Math 6	Basic math	One school year	128	General	Middle	Quasi-experimental	Before 2003	No	Nonjournal	0.42	
Koedinger et al. (1997)	Cognitive Tutor Algebra—PUMP	Algebra	One school year	225	General	High	Quasi-experimental	Before 2003	No	Journal	0.49	
Morgan & Ritter (2002)	Cognitive Tutor Algebra I	Algebra	One school year	384	General	High	True experimental	Before 2003	No	Nonjournal	0.23	
Pane et al. (2010)	Cognitive Tutor Geometry	Geometry	One school year	699	General	High	True experimental	2006–2010	No	Journal		–0.19
Plano, Ramey, & Achilles (2007)	Cognitive Tutor Algebra	Algebra	One school year	779	Low achievers	High	Quasi-experimental	2003–2005	No	Nonjournal	–0.66 ^d	–0.48
Ritter et al. (2007)	Cognitive Tutor Algebra I	Algebra	One school year ^e	342	General	High	True experimental	ng	No	Journal	0.29	
Sarkis (2004)	Cognitive Tutor Algebra I	Algebra	One school year	4,649	General	High	Quasi-experimental	2003–2005	No	Nonjournal	0.13	
Shneyderman (2001)	Cognitive Tutor Algebra I	Algebra	One school year	663	General	High	Quasi-experimental	Before 2003	No	Nonjournal	0.14	
Smith (2001)	Cognitive Tutor Algebra	Algebra	More than one school year	445	Low achievers	High	True experimental	Before 2003	No	Nonjournal		–0.07
Stankov et al. (2008) (1)	Tutor-Expert System	Basic math	Short term	18	General	Elementary	Quasi-experimental	2006–2010	No	Journal	0.92	1.05
Stankov et al. (2008) (2)	Tutor-Expert System	Basic math	Short term	18	General	Elementary	Quasi-experimental	2006–2010	No	Journal	0.08	0.11
Stankov et al. (2008) (3)	Tutor-Expert System	Basic math	Short term	48	General	Elementary	Quasi-experimental	2006–2010	No	Journal	0.00	0.31
Wallis (2005)	Wayang Outpost	Basic math	Short term	218	General	High	True experimental	2003–2005	No	Nonjournal	–0.25	

Note. ITS = intelligent tutoring system; ES = effect size; ng = not given.

^a The sample sizes reported in this table are the total sample sizes of each independent study. Grubbs (1950) tests showed that among the total sample sizes, five of them were detected as outliers. They are 17,164, 4,649, 3,566, 3,136, and 1,404, for which the nearest neighbor is 799. ^b These are unadjusted overall effect sizes for each independent sample. ^c These are adjusted overall effect sizes for each independent sample. ^d The original effect size extracted from this study was –1.57. It was detected as an outlier in Grubbs (1950) tests. In the analyses, we reset the value to –0.66, its next nearest neighbor among unadjusted overall effect sizes. ^e The Ritter et al. (2007) study reported an outcome measure after one semester of the intervention, and it also reported two outcome measures after one school year of the intervention. The study sample remained same. So we pooled the outcome measures and extracted one overall effect size for this study. However, we categorized the ITS duration as one school year.

ITS with human tutoring. We narratively reported the results of the studies that compared ITS with human tutors or home work conditions later in this section. We did not include them in the analyses described below so as to have a single clear comparison group. Therefore, the 61 effect sizes of ITS in comparison to regular classroom instruction made up the data for the results that follow.

With the 61 effect sizes, we formed three different data sets. The first data set included unadjusted overall effect sizes. It consisted of 26 effect sizes with each independent sample contributing one overall effect size to the data set. Here, if multiple effect sizes were extracted from the same sample, these effect sizes were averaged to estimate the overall effectiveness of ITS on this independent sample. The second data set included all unadjusted effect sizes. This data set consisted of all 44 unadjusted effect sizes from the 26 independent samples. The third data set consisted of 17 adjusted overall effect sizes from 17 independent samples. Some independent samples provided both an unadjusted and an adjusted overall effect size, whereas some only provided one type of overall effect sizes or the other.

We conducted analyses on adjusted and unadjusted effect sizes separately. One may argue that it would be beneficial to pool the two types of effect sizes so that the analyses would include all of the 31 effect sizes from the 31 studies. However, we think the benefits of conducting the analyses separately outweigh those of analyzing them together. We have three justifications. First, distinguishing adjusted and unadjusted effect sizes would allow us to examine whether estimates of ITS effectiveness differs with or without controlling for confounding factors. Second, the number of studies in the analyses did not increase significantly even if we analyze the effect sizes together. Specifically, if the effect sizes were analyzed together, the total number of studies would be increased from 26 (i.e., the number of unadjusted effect sizes) to 31 (i.e., the total number of independent studies or samples). Finally, analyzing the two types of effect sizes separately helps in interpretation. Differentiating adjusted and unadjusted effect sizes and integrating them separately allows us to provide clearer information regarding what each estimate of effect refers to with regard to the achievement outcome.

We conducted Grubbs (1950) tests to look for statistical outliers before calculating the average effect sizes. The Grubbs tests showed that, among the unadjusted overall effect sizes ($k = 26$), one effect size ($g = -1.57$) appeared to be an outlier (i.e., Plano, Ramey, & Achilles, 2007). We found that the Plano et al. (2007) study provided information for both an adjusted (adjusted by pretest scores, $g = -0.48$) and an unadjusted effect size ($g = -1.57$). Clearly then, the unadjusted effect size was strongly impacted by the preexisting differences between the treatment and comparison groups. We reset the effect size to -0.66 , its next nearest neighbor among the unadjusted overall effect sizes. Among all the unadjusted effects sizes ($k = 44$), the effect size ($g = -1.57$) from the Plano et al. (2007) study again appeared to be an outlier. We reset the effect size to -1.03 , its next nearest neighbor. The Grubbs tests detected no outliers among the adjusted overall effect sizes. We also conducted Grubbs tests on ITS and comparison group sample sizes. Again, we reset the outlier sample sizes to their nearest neighbors. We conducted analyses after adjusting the outlier sample sizes.²

As Table 1 shows, 10 different ITS were studied in the 31 independent studies comparing ITS with regular classroom instruction. Cognitive Tutor by Carnegie Learning was the most frequently studied. Specifically, Cognitive Tutor for algebra learning was evaluated in 16 studies; Cognitive Tutor for math was studied in three studies; in one study, Cognitive Tutor was used for geometry. As mentioned in the introduction, the WWC had completed four reviews on the effectiveness of Cognitive Tutor. Four other ITS (i.e., Larson Pre-Algebra/Algebra I, Achieve Now, iLearn Math, and Plato Algebra) were evaluated in a national-level study (see Campuzano et al., 2009; Dynarski et al., 2007) and were also reviewed by the WWC. The other two ITS that were relatively frequently studied were Wayang Outpost (see Arroyo et al., 2010; Beal et al., 2010; Walles, 2005) and Tutor-Expert System (see the three studies by Stankov, Rosic, Zitko, & Grubisic, 2008). Online Remediation Software appeared in two studies by Biesinger and Crippen (2008). AnimalWatch (Beal, Arroyo, Cohen, & Woolf, 2007) and Intelligent Tutoring, Evaluation and Diagnosis (Hwang et al., 2008) each appeared once.

Because Cognitive Tutor was most frequently studied, we briefly describe its mechanism, scope of use, and the length of its implementation. Cognitive Tutor is built on a cognitive theory called adaptive control of thought (Anderson et al., 1995). Cognitive Tutor presents students with a series of problems and adaptively identifies a student's problem-solving strategy through his or her actions and comparisons with correct solution approaches and misconceptions generated by the program's cognitive model, a process called *model tracing*. Five curricula have been developed with Cognitive Tutor as their software component and have been used by more than 500,000 students in approximately 2,600 schools across the United States as of 2010 (WWC, 2010a). They are Bridge to Algebra, Algebra I, Geometry, Algebra II, and Integrated Math. In these curricula, students generally spend three class periods per week in regular classroom learning and two class periods in computer lab using Cognitive Tutor. In most of the evaluation studies included in this meta-analysis, students used Cognitive Tutor for one school year or one semester.

Overall Effectiveness of ITS on Students' Mathematical Learning

We conducted meta-analyses on the data sets of unadjusted and adjusted overall effect sizes to examine the overall effectiveness of ITS on students' mathematical learning, compared with that of regular classroom instruction. All effect sizes were weighted by inverse variances. Of the 26 unadjusted overall effect sizes, 17 were in a positive direction, eight were in a negative direction, and one was exactly 0. Under a fixed-effect model, the average effect size was 0.05, 95% CI [.02, .09], $p = .005$, and was significantly different from 0. Under a random-effects model, the average effect size was 0.09, 95% CI [-.03, .20], $p = .136$, and was not significantly from 0. There was a high degree of heterogeneity

² It is worth mentioning that we also calculated the average effect sizes and ran moderator analyses on the effect sizes without adjusting the outlier sample sizes. We found very minor differences in the analysis results between those with and without adjusted outlier sample sizes. These differences were not sufficient to lead to any major changes in conclusions. Therefore, we choose to only report the analysis results with the sample size outliers adjusted.

among the 26 unadjusted overall effect sizes, $Q_t(25) = 180.80$, $p = .000$. This indicates that it was unlikely that sampling error alone was responsible for the variance among the effect sizes; instead, some other factors likely played a role in creating variability as well.

Of the 17 adjusted overall effect sizes, 10 were in a positive direction and seven were in a negative direction. Under a fixed-effect model, the average effect size was 0.01, 95% CI $[-.04, .06]$, $p = .792$, and was not significantly different from 0. Under a random-effects model, the average effect size also was 0.01, 95% CI $[-.10, .12]$, $p = .829$, and was also not statistically significantly different from 0. There was a high degree of heterogeneity among the 17 adjusted overall effect sizes, $Q_t(16) = 54.01$, $p = .000$. Again, this suggested that it was unlikely that sampling error alone was responsible for the total variance among the effect sizes.

Examining Publication Bias

We conducted Duval and Tweedie's (2000) trim and fill procedure to assess the possible effects of publication bias. For unadjusted overall effect sizes, there was evidence that three studies on the left side of the distribution might have been missing under both a fixed-effect model and a random-effects model. The overall average effect size after imputing the three additional values was 0.04 under a fixed-effect model and was 0.03 under a random-effects model. The average effect size for the observed effect sizes, as reported previously, was 0.05 under a fixed-effect model and 0.09 under a random-effects model. This implies that the average effect of ITS might have been slightly overestimated.

For adjusted overall effect sizes, three studies on the left side of the distribution were projected as missing under a fixed-effect model, and six effect sizes on the left side of the mean were projected as missing under a random-effects model. The overall average effect sizes after imputing the three additional values ranged from -0.04 to -0.01 using a fixed-effect model; the overall average effect size after imputing the six additional values was -0.09 using a random-effects model. The average of the observed effect sizes, as reported previously, were 0.01 under both a fixed-effect model and random-effects model. Therefore, there was little evidence that publication bias had much impact on the average effect size in this case.

Testing for Moderators on the Unadjusted and Adjusted Overall Effect Sizes

We conducted moderator analyses exploring nine variables that could possibly have an impact on ITS's effects. We chose these nine variables for two reasons. First, they represented important features of ITS intervention or research methodology. Second, there were at least two effect sizes associated with each of the category of the variable, in the data sets of both unadjusted and adjusted overall effect sizes, to allow meaningful analyses.³

Tables 2 and 3 present the results of testing for moderators on the unadjusted and adjusted overall effect sizes, respectively. In each data set, the number of effect sizes involved might be different. For variables with more than two categories, we first conducted comparisons using all of the categories and then regrouped them to create a two-group comparison. We included the results of further analyses on the two-group comparison in the tables as well.

Subject matter. Testing results showed that the effectiveness of ITS did not differ for different subject matters under a fixed-effect model, $Q_b(1) = .12$, $p = .726$, nor did it differ under a random-effects model, $Q_b(1) = .62$, $p = .431$, for unadjusted effect sizes. The advantage of using ITS, compared with regular classroom instruction, was significant only for basic math under the fixed-effect model, indicated by the fact that the confidence interval of the effect size ($g = .06$) did not contain 0.

For adjusted effect sizes, results showed the effectiveness of ITS on students' learning of basic math appeared to be greater than that of learning algebra under a fixed-effect model, $Q_b(1) = 9.10$, $p = .003$, but not under a random-effects model, $Q_b(1) = 1.62$, $p = .204$.⁴ Specifically, under a fixed-effect model, the average effectiveness of ITS was $g = 0.11$, 95% CI $[.04, .19]$, on helping students learn basic math and $g = -0.05$, 95% CI $[-.13, .02]$, on learning algebra.

ITS duration. For unadjusted effect sizes, the effectiveness of ITS differed depending on the length of instruction under both a fixed-effect model, $Q_b(2) = 16.28$, $p = .000$, and a random-effects model, $Q_b(2) = 6.42$, $p = .04$. Further analyses revealed no difference between the short-term and one-semester ITS interventions. We therefore compared the combination of short-term and one-semester interventions with interventions of one school year or longer. We found that under a fixed-effect model, the average effectiveness of ITS was greater when the interventions lasted for less than one school year, $g = 0.23$, 95% CI $[.13, .32]$, than when they lasted for one school year or longer, $g = .02$, 95% CI $[-.02, .06]$; under a random-effects model, the average effectiveness of ITS was also greater when the interventions lasted for less than one school year, $g = 0.26$, 95% CI $[.08, .44]$, than that of when they lasted for one school year or longer, $g = -.01$, 95% CI $[-.15, .14]$.

For adjusted effect sizes, results showed that ITS effectiveness also differed depending on the duration of intervention under both a fixed-effect model, $Q_b(2) = 13.88$, $p = .001$, and a random-effects model, $Q_b(2) = 14.71$, $p = .001$. Further analyses revealed that the effects differed depending on whether the ITS intervention lasted for one school (or longer) or less than one school year under both a fixed-effect model, $Q_b(1) = 6.48$, $p = .011$, and a random-effects model, $Q_b(1) = 7.40$, $p = .007$.

Sample achievement level. We categorized study samples in terms of the academic achievement level of the subjects, using the way they were reported in the primary studies to categorize samples. Two types of student samples appeared. One consists of general students, a population that includes students of all achievement levels. Another consists of low achievers. There were three studies that reported results for low achievers (i.e., Biesinger & Crippen, 2008; Plano et al., 2007; Smith, 2001). For unadjusted effect sizes, under a fixed-effect model, the effectiveness of ITS on

³ The second reason led us to drop a number of variables we initially intended to study. For example, we hoped to compare whether there was a difference in the effects of ITS when they were used to substitute for regular classroom instruction and when they were used only as a supplement to regular classroom instruction. We were unable to do so because, for the 17 adjusted effect sizes, only one effect size was associated with ITS used as substitute, versus 16 effect sizes associated with ITS as a supplement.

⁴ For adjusted overall effect sizes, we dropped one effect size associated with geometry, $g = -.19$, 95% CI $[-.34, -.04]$

Table 2
Testing for Moderators of the Unadjusted Effect Sizes

Variable	k	Fixed			Random		
		g	95% CI	Q _b	p _b	g	95% CI
Subject							
Basic math	12	.06	[.01, .09]	.12	.726	.14	[-.03, .32]
Algebra	14	.05	[-.01, .14]			.05	[-.11, .21]
ITS duration				16.28***	.000		6.42*
One school year or longer	15	.02	[-.02, .06]			-.01	[-.15, .14]
One semester	4	.24	[.13, .36]			.28	[.10, .45]
Short term	7	.20	[.04, .36]			.23	[-.13, .58]
ITS duration (further analysis)				16.07***	.000		5.10*
One school year or longer	15	.02	[-.02, .06]			-.01	[-.15, .14]
Less than one school year	11	.23	[.13, .32]			.26	[.08, .44]
Sample achievement level				46.13***	.000		.60
General students	24	.09	[.05, .12]			.12	[.02, .21]
Low achievers	2	-.42	[-.55, -.28]			-.23	[-.1.08, .63]
Schooling level				2.11	.349		.58
Elementary school	3	.21	[-.22, .62]			.25	[-.28, .77]
Middle school	6	-.001	[-.09, .09]			.02	[-.24, .29]
High school	14	.06	[.02, .11]			.09	[-.07, .25]
Schooling level (further analysis)				.51	.473		.37
Elementary school	3	.21	[-.22, .62]			.25	[-.28, .77]
Secondary school	23	.05	[.01, .09]			.08	[-.04, .20]
Sample size				14.28**	.001		.70
Less than 200	9	.27	[.09, .45]			.21	[-.14, .56]
Over 200 but less than 1,000	12	-.02	[-.08, .03]			.05	[-.14, .25]
Over 1,000	5	.09	[.04, .13]			.06	[-.09, .20]
Sample size (further analysis)				5.94*	.015		.74
Less than 200	9	.27	[.09, .45]			.21	[-.14, .56]
Over 200	17	.04	[.01, .08]			.05	[-.08, .17]
Research design				7.97**	.005		.37
Quasi-experimental	15	.09	[.05, .14]			.12	[-.07, .31]
True experimental	11	-.01	[-.07, .05]			.05	[-.09, .19]
Year of data collection				10.51**	.005		3.01
Before 2003	7	.20	[.09, .30]			.18	[-.02, .39]
Between 2003-2005	7	.02	[-.02, .07]			-.08	[-.30, .14]
Between 2006-2010	8	-.01	[-.09, .07]			.05	[-.14, .24]
Year of data collection (further analysis)				1.69	.193		.02
Before 2006	14	.05	[.01, .09]			.03	[-.13, .19]
2006 and after	8	-.01	[-.09, .07]			.05	[-.14, .24]
Counterbalanced testing				10.43**	.001		1.09
No	20	.09	[.05, .13]			.13	[-.02, .27]
Yes	5	-.05	[-.12, .02]			-.002	[-.20, .19]
Report type				24.45***	.000		10.03**
Peer-reviewed journal	10	.28	[.18, .37]			.30	[.17, .43]
Nonjournal	16	.02	[-.02, .05]			-.01	[-.15, .13]
Measurement timing ^a				18.81***	.000		5.71
End of school year	18	.01	[-.03, .04]			.01	[-.14, .15]
End of semester	3	.29	[.15, .43]			.34	[.11, .57]
Immediately after intervention	6	.19	[.02, .35]			.16	[-.29, .60]

Table 2 (continued)

Variable	<i>k</i>	Fixed			Random		
		<i>g</i>	95% CI	Q_b	<i>p_b</i>	<i>g</i>	95% CI
Measurement timing (further analysis)	18	.01	[-.03, .04]	17.84***	.000	.01	[-.14, .15]
End of school year	9	.25	[-.14, .35]			.26	[-.01, .51]
Outcome type ^b	3	.29	[-.15, .43]	16.57**	.002	.33	[-.10, .56]
Course grades	2	.14	[-.01, .16]			.22	[-.12, .55]
Course passing rates	11	.13	[-.02, .27]			.10	[-.24, .45]
Specifically designed tests	6	.05	[-.05, .16]			.04	[-.21, .28]
Modified standardized tests	14	.02	[-.02, .05]			.03	[-.13, .19]
Standardized tests				13.19***	.000		
Outcome type (further analysis)	16	.19	[-.11, .27]			.20	[-.01, .39]
Course-related outcome measures	20	.02	[-.02, .06]			.03	[-.11, .16]
Standardized test measures							

Note. CI = confidence interval; Q_b denotes the heterogeneity status between all subcategories of a particular variable under testing, with degrees of freedom equal to moderator levels minus one; ITS = intelligent tutoring system.

^a Testing for moderators was conducted on the all unadjusted effect sizes so that total number of *k* exceeded 26. ^b Testing for moderators was conducted on the all unadjusted effect sizes so that total number of *k* exceeded 26.

* $p < .05$. ** $p < .01$. *** $p < .001$.

helping general students learn mathematical subjects, $g = .09$, 95% CI [.05, .12], was greater than on helping low achievers, $g = -.42$, 95% CI [-.55, -.28], $Q_b(1) = 46.13$, $p = .000$. The difference was not significant under a random-effects model, $Q_b(1) = 0.60$, $p = .438$. Overall, ITS appeared to have a positive impact on general students. For low achievers, the average effect was negative under both fixed-effect and random-effects models.

For adjusted effect sizes, the effects of ITS on helping general students learn mathematical subjects, $g = .04$, 95% CI [-.02, .09], were greater than on helping low achievers, $g = -.18$, 95% CI [-.32, -.05], $Q_b(1) = 9.24$, $p = .002$, under a fixed-effect model. The difference was not significant under a random-effects model, $Q_b(1) = 1.31$, $p = .253$. In this analysis, the only average effect size that was significantly different from 0 was the negative effect indicating that regular classroom instruction compared favorably with ITS under a fixed-effect model.

Schooling level. The unadjusted overall effect sizes were associated with samples of three schooling levels: (a) elementary school, which included K–5 grade levels; (b) middle school, which included Grades 6–8; and (c) high school, which included Grades 9–12.⁵ The relative effectiveness of ITS on students' mathematical learning did not vary significantly in terms of schooling level under either a fixed-effect model, $Q_b(2) = 2.11$, $p = .349$, or a random-effects model, $Q_b(2) = 0.58$, $p = .749$. We regrouped the effect sizes into elementary school and secondary school levels. Again, results showed that the difference was not significant under either a fixed-effect model, $Q_b(1) = 0.51$, $p = .473$, or a random-effects model, $Q_b(1) = 0.37$, $p = .545$.

For the adjusted overall effect sizes, the effectiveness of ITS varied significantly in terms of schooling level under a fixed-effect model, $Q_b(2) = 14.29$, $p = .001$, but not under a random-effects model, $Q_b(2) = 3.07$, $p = .215$. The average effect sizes suggested that the effects of ITS might be most pronounced for students in elementary school, $g = .41$, 95% CI [-.01, .84], compared with those in middle school, $g = .09$, 95% CI [.01, .17] and in high school, $g = -.09$, 95% CI [-.17, -.02]. However, when the effect sizes were regrouped into elementary and secondary school levels, no statistically significant difference was found between them under either a fixed-effect model, $Q_b(1) = 3.61$, $p = .057$, or a random-effects model, $Q_b(1) = 3.19$, $p = .074$.

Sample size. Among the 26 unadjusted overall effect sizes, nine were associated with sample sizes less than 200, 12 were associated with sample sizes over 200 but less than 1,000, and five were associated with sample sizes over 1,000. The unadjusted effectiveness of ITS corresponding to each of these three sample size categories varied significantly under a fixed-effect model, $Q_b(2) = 14.28$, $p = .001$, but not under a random-effects model, $Q_b(2) = 0.70$, $p = .704$. Further analyses revealed that the effects were greater when the study sample sizes were less than 200 than when the sample sizes were over 200 under a fixed-effect model, $Q_b(2) = 5.94$, $p = .015$, but not under a random-effects model, $Q_b(2) = 0.74$, $p = .389$.

⁵ We did not include three unadjusted and two adjusted effect sizes associated with studies in which samples were across both middle school and high school. It is also worthy to note that there were only three studies involved elementary school students and all of them were conducted by a same research team.

Table 3
Testing for Moderators of the Adjusted Overall Effect Sizes

Variable	<i>k</i>	Fixed				Random			
		<i>g</i>	95% CI	Q_b	p_b	<i>g</i>	95% CI	Q_b	p_b
Subject				9.10**	.003			1.62	.204
Basic math	9	.11	[.04, .19]			.11	[−.05, .28]		
Algebra	7	−.05	[−.13, .02]			−.03	[−.19, .12]		
ITS duration				13.88**	.001			14.71**	.001
One school year or longer ^a	10	−.02	[−.07, .04]			−.08	[−.20, .05]		
One semester	2	.06	[−.13, .24]			.06	[−.13, .24]		
Short term	5	.52	[.24, .79]			.52	[.24, .79]		
ITS duration (further analysis)				6.48*	.011			7.40**	.007
One school year of longer	10	−.02	[−.07, .04]			−.08	[−.20, .05]		
Less than one school year	7	.19	[.04, .34]			.29	[.06, .53]		
Sample achievement level				9.24**	.002			1.31	.253
General students	14	.04	[−.02, .09]			.05	[−.06, .16]		
Low achievers	3	−.18	[−.32, −.05]			−.16	[−.49, .18]		
Schooling level				14.29**	.001			3.07	.215
Elementary school	3	.41	[−.01, .84]			.42	[−.04, .89]		
Middle school	4	.09	[.01, .17]			−.004	[−.20, .19]		
High school	8	−.09	[−.17, −.02]			−.01	[−.19, .16]		
Schooling level (further analysis)				3.61	.057			3.19	.074
Elementary school	3	.41	[−.01, .84]			.42	[−.04, .89]		
Secondary school	14	.001	[−.05, .05]			−.01	[−.13, .10]		
Sample size				12.49**	.002			2.77	.251
Less than 200	6	.18	[−.06, .43]			.24	[−.11, .59]		
Over 200 but less than 1000	8	−.08	[−.16, .01]			−.06	[−.20, .09]		
Over than 1,000	3	.09	[.01, .16]			.06	[−.10, .22]		
Sample size (further analysis)				2.03	.154			1.96	.162
Less than 200	6	.18	[−.06, .43]			.24	[−.11, .59]		
Over 200	11	−.001	[−.05, .05]			−.02	[−.14, .09]		
Research design				.92	.338			1.09	.296
Quasi-experimental	9	−.06	[−.20, .08]			.17	[−.16, .50]		
True experimental	8	.02	[−.04, .07]			−.01	[−.11, .08]		
Year of data collection				2.31	.315			.715	.699
Before 2003	3	−.08	[−.25, .08]			−.08	[−.25, .08]		
Between 2003–2005	4	.03	[−.04, .10]			−.07	[−.33, .19]		
Between 2006–2010	8	−.03	[−.12, .05]			.00	[−.13, .14]		
Year of data collection (further analysis)				.79	.375			.53	.468
Before 2006	7	.01	[−.05, .08]			−.08	[−.26, .10]		
2006 and after	8	−.03	[−.12, .05]			.004	[−.13, .14]		
Counterbalanced testing				8.89**	.003			.44	.507
No	12	−.08	[−.16, −.004]			−.01	[−.17, .14]		
Yes	5	.08	[.01, .14]			.06	[−.09, .21]		
Report type				.69	.407			1.98	.160
Peer-reviewed journal	6	−.04	[−.17, .08]			.23	[−.11, .57]		
Nonjournal	11	.02	[−.04, .07]			−.03	[−.15, .09]		

Note. CI = confidence interval; Q_b denotes the heterogeneity status between all subcategories of a particular variable under testing; ITS = intelligent tutoring system.

^a This subcategory included one study in which the ITS intervention lasted more than one school year.

* $p < .05$. ** $p < .01$.

The adjusted effectiveness of ITS associated with each of these three sample size categories varied significantly under a fixed-effect model, $Q_b(2) = 12.49$, $p = .002$, but not under a random-effects model, $Q_b(2) = 2.77$, $p = .251$. Further analyses showed that the effects did not differ significantly between studies with sample sizes of less than 200 and those with sample sizes over 200 under a fixed-effect model, $Q_b(1) = 2.03$, $p = .154$, nor did it under a random-effects model, $Q_b(1) = 1.96$, $p = .162$.

Research design. For unadjusted overall effect sizes, 15 were from quasi-experimental studies and 11 were from true experiments. The average of the effect sizes from the quasi-experiments, $g = .09$, 95% CI [.05, .14], was larger than that of from true

experiments, $g = −.01$, 95% CI [−.07, .05], under a fixed-effect model, $Q_b(1) = 7.97$, $p = .005$. Only the average effect for studies using quasi-experimental designs was significantly different from 0. The difference was not significant under a random-effects model, $Q_b(1) = 0.37$, $p = .544$.

The average of adjusted overall effect sizes from quasi-experiments and true experiments did not differ under either a fixed-effect model, $Q_b(1) = 0.92$, $p = .338$, or under a random-effects model, $Q_b(1) = 1.09$, $p = .296$. None of the average effects were significantly different from 0.

Year of data collection. For unadjusted overall effect sizes, the effects varied depending on the year in which the data were

collected under a fixed-effect model, $Q_b(2) = 10.51, p = .005$, but not under a random-effects model, $Q_b(2) = 3.01, p = .222$. Only the average effect for studies conducted before 2003 (showing a positive ITS effect) appeared significantly different from 0 under a fixed-effect model. For adjusted overall effect sizes, the effects did not differ significantly in terms of data collection time under either a fixed-effect model, $Q_b(2) = 2.31, p = .315$, or a random-effects model, $Q_b(2) = 0.715, p = .699$.

Counterbalanced testing. For unadjusted overall effect sizes, the impact of ITS on students' mathematical learning appeared to be lower in studies with counterbalanced testing, $g = -.05$, 95% CI $[-.12, .02]$, than in studies without counterbalanced testing, $g = .09$, 95% CI $[.05, .13]$, under a fixed-effect model, $Q_b(1) = 10.43, p = .001$. The average effect size from counterbalanced studies did not differ from 0, whereas the average effect from studies not using counterbalancing did. The difference was not significant under a random-effects model, $Q_b(1) = 1.09, p = .297$.

For adjusted overall effect sizes, the impact of ITS appeared to be significantly larger in studies with counterbalanced testing, $g = .08$, 95% CI $[.01, .14]$, than that of studies without counterbalanced testing, $g = -.08$, 95% CI $[-.16, -.004]$ under a fixed-effect model, $Q_b(1) = 8.89, p = .003$. The difference was not statistically significant under a random-effects model, $Q_b(1) = 0.44, p = .507$.

Report type. We grouped the reports into two categories: peer-reviewed journal reports and nonjournal reports. Nonjournal reports include government reports, conference papers, private reports, master's theses, and doctoral dissertations. For unadjusted overall effect sizes, the average effect size in peer-reviewed journals, $g = .28$, 95% CI $[.18, .37]$, was higher than that of nonjournal reports, $g = .02$, 95% CI $[-.02, .05]$, under a fixed-effect model, $Q_b(1) = 24.45, p = .000$. The average effect of ITS was positive in studies in peer-reviewed journals. Under a random-effects model, the average effect size in peer-reviewed journals, $g = .30$, 95% CI $[.17, .43]$, was also statistically significantly higher than that of nonjournal reports, $g = -.01$, 95% CI $[-.15, .13]$, $Q_b(1) = 10.03, p = .002$; again, the average effect of ITS was positive in studies in peer-reviewed journals. For adjusted overall effect sizes, the average effect size in peer-reviewed journals was not different from that of nonjournal reports under a fixed-effect model, $Q_b(1) = 0.69, p = .407$, nor was the case under a random-effect model, $Q_b(1) = 1.98, p = .160$.

Testing for Moderators on All Unadjusted Effect Sizes

The data set of all unadjusted effect sizes consisted of all of the 44 unadjusted effect sizes, not averaged within independent samples. This data set allowed us to study two moderators: the measurement timing and the types of outcomes. The analysis results are also included in Table 2.

Measurement timing. Within a single independent sample, we averaged the effects sizes that related to the same measurement timing. This reduced the 44 unadjusted effect sizes to 27 unadjusted effect sizes that were either associated with outcomes measured at the end of the school year or measured sooner than the end of the school year. The effectiveness of ITS when measured at the end of school year, $g = .01$, 95% CI $[-.03, .04]$, was lower than that when measured sooner than that, $g = .25$, 95% CI $[.14, .35]$, under a fixed-effect model, $Q_b(1) = 17.84, p = .000$, but not under a random-effects model, $Q_b(1) = 3.02, p = .082$.

Outcome type. As above, we averaged the effect sizes corresponding to the same outcome type within each independent sample. This resulted in 36 effect sizes in our analyses. Five different types of outcomes appeared in the studies. They are (a) course grades, (b) course passing rates, (c) scores from tests developed by teachers or researchers to specifically measure students' learning on the knowledge content that was covered by interventions, (d) scores from modified standardized tests that were either substrands of standardized tests or tests made up of some of the released standardized test questions, and (e) scores from standardized tests. Our preliminary analyses showed that there was no statistically significant difference between the average effect size associated with course grades or course passing rates and that of specifically designed tests, nor was a statistically significant difference between the average effect size associated with modified standardized tests and that of standardized tests. Thus, we grouped the first three outcome types into course-related measures and the last two outcome types into measures from standardized tests. Results show that under a fixed-effect model, the average effect size for course-related measures was $g = 0.19$, 95% CI $[.11, .27]$, and $g = 0.02$, 95% CI $[-.02, .06]$ for measures from standardized tests, $Q_b(1) = 13.19, p = .000$. The difference was not significant under a random-effects model, $Q_b(1) = 2.09, p = .148$. Course-related measures showed a larger and positive ITS effect and were significantly different from 0 under both models.

Effectiveness of ITS in Comparison to Other Treatment Conditions

In comparison to homework. Mendicino et al. (2009) compared 28 fifth-grade students learning math in two different homework conditions over a period of 1 week. In one condition, students completed paper-and-pencil homework. In another condition, students completed Web-based homework using the ASSISTment system. ASSISTment is a Web-based homework system that facilitates students' learning by providing scaffolds and hints. A number sense problem set and a mixed-problem test were used to measure students' learning after each homework condition. To reduce the possibility that other factors might impact learning, Mendicino et al. implemented counterbalanced procedures so that all students participated in both paper-and-pencil and Web-based conditions. They were tested both before and after the intervention. Mendicino et al. reported an adjusted effect size of 0.61 favoring ITS and concluded that students learned significantly more with the help of the Web-based system than by working on paper-and-pencil homework.

Radwan (1997) compared the math performance of 52 fourth graders. Half were tutored using the Intelligent Tutoring System Model and the other half received no tutoring but worked on completing homework, both during the fifth period of school days. The experiment lasted for a total of 15 hr 40 min every day for 4 weeks. Students' learning was measured through a pretest and posttest of the Computerized Achievement Tests. Radwan's t tests on the test scores concluded that the experimental group performed significantly better than the control group did. On the basis of the overall score on the Computerized Achievement Tests, we found that this study yielded an unadjusted $g = .40$, $SE = .28$, and an adjusted $g = .60$, $SE = .28$.

In comparison to human tutoring. Beal et al. (2010) studied the effectiveness of AnimalWatch, an intelligent tutoring system designed to help students learn basic computation and fraction skills to enhance problem-solving performance. The participants were sixth graders enrolled in a summer academic skills class in Los Angeles, California. Once per week for 4 weeks, 13 sixth graders spent 1 hr with math tutors and then 45 min with AnimalWatch, and 12 sixth graders learned math with their tutors (each tutor helped four to six students) using small group activities including blackboard lessons and worksheet practice. The mean proportion of correct scores was used to measure students' performance. Beal et al. concluded that students who spent half of their time using ITS and half of time with human tutors improved as much as did those who spent the entire time learning with a human tutor. We found that this study yielded an unadjusted $g = .20$, $SE = .39$.

Discussion

Summary of the Evidence

Findings of this meta-analysis suggest that, overall, ITS had no negative and perhaps a very small positive effect on K–12 students' mathematical learning relative to regular classroom instruction. When the effectiveness was measured by posttest outcomes and without taking into account the potential influence of other factors, the average unadjusted effect size was .05 under a fixed-effect model and .09 under a random-effects model favoring ITS over regular classroom instruction. After controlling for the influence of other variables (e.g., pretest scores), the average adjusted effect size was .01 under both a fixed-effect model and random-effects model also favoring the ITS condition. However, the average relative effectiveness of ITS did not appear to be significantly different from 0 except when effect sizes were unadjusted and a fixed-effect analysis model was used. Also, whether controlling for other factors or not, there was a high degree of heterogeneity among the effect sizes.

Very few studies compared ITS with homework or human tutoring. The few existing studies showed that when compared

with homework or human tutoring, the relative effectiveness of ITS appeared to be small to modest, with effect sizes ranging from .20 to .60.

Testing for moderators yielded some informative findings. Table 4 presents a summary of the findings from moderator analyses using two different estimates of effect (i.e., unadjusted and adjusted effect sizes) and two analysis models (i.e., fixed-effect and random-effects models). Two findings were relatively robust. First, the effects appeared to be greater when the ITS intervention lasted for less than a school year than when it lasted for one school year or longer. This effect appeared regardless of whether the moderator analyses were conducted on unadjusted or adjusted effect sizes with a fixed-effect or random-effects model. Second, the effects of ITS appeared to be greater when the study samples were general students than when the samples were low achievers. And under a fixed-effect model, this difference was statistically significant regardless of whether the analyses were conducted on unadjusted or adjusted effect sizes.

Also, there was some evidence for the following three findings related to the methodology of the study: (a) The effectiveness of ITS appeared to be largest when the learning outcomes were measured before the end of the school year, (b) the effects of ITS appeared to be greater when measured by course-related outcomes than when measured by standardized tests, and (c) the average effect size of studies with smaller sample sizes appeared to be bigger than that of larger sample sizes. In general, these results are consistent with the findings related to methodological characteristics of primary studies in numerous meta-analyses.

Overall Effectiveness of ITS

The conclusion that ITS had no negative and perhaps a very small positive effect on K–12 students' mathematical learning relative to regular classroom instruction is largely congruent with the WWC's conclusions regarding the effects of math educational software programs (WWC, 2004, 2010a, 2010b). Specifically, the WWC (2010a) concluded that Carnegie Learning Curricula and Cognitive Tutor software had no discernible effects on mathematics achievement for high school students but Cognitive Tutor[®]

Table 4

Findings From Testing for Moderators Across Two Types of Effect Sizes and Two Analysis Models

Variable	ITS favored for	Unadjusted		Adjusted	
		Fixed	Random	Fixed	Random
Subject	Basic math	Yes	Yes	Yes+	Yes
ITS duration	Less than one school year	Yes+	Yes+	Yes+	Yes+
Sample achievement level	General students	Yes+	Yes	Yes+	Yes
Schooling level	Elementary school	Yes	Yes	Yes	Yes
Sample size	Sample size less than 200	Yes+	Yes	Yes	Yes
Research design	Quasi-experiments	Yes+	Yes	No	Yes
Year of data collection	Before 2006	Yes	No	Yes	No
Counterbalanced testing	No	Yes+	Yes	No+	No
Report type	Peer-reviewed journal	Yes+	Yes+	No	Yes
Measurement timing	Sooner than end of school year	Yes+	Yes		
Outcome type	Course-related outcome measures	Yes+	Yes		

Note. Yes denotes that the subcategory, for example, basic math, appears to be favored over the other subcategory or subcategories of that variable (i.e., subject). A + denotes that the effects of the intelligent tutoring system (ITS) on the favored feature (i.e., subcategory), for example, basic math, was statistically significantly greater than those on the other feature (i.e., subcategory), such as algebra.

Algebra I had potentially positive effects on ninth graders' math achievement. The WWC (2004) found that students who used Cognitive Tutor earned significantly higher scores on the Educational Testing Service Algebra I test and on their end-of-semester grades than their counterparts who were taught with traditional instruction. Furthermore, the WWC (2010b) concluded that PLATO Achieve Now had no discernible effects for six graders' math achievement but the WWC considered the extent of evidence to be small.

It is relevant to mention that the WWC conclusions were based on a very limited number of studies that met their evidence standards or met their standards with reservation. The present meta-analysis included all seven reports that had been identified as meeting the WWC's evidence standards or meeting their standards with reservation. This meta-analysis covered studies of many other ITS programs in addition to the two reviewed by the WWC. Thus, despite the differences in review scopes and methodology, the finding that ITS appeared to have no negative and perhaps a small positive effect on students' mathematical achievement is largely consistent with the conclusion from the WWC reviews.

Comparing the present meta-analysis with a recent meta-review by VanLehn (2011) illuminates an interplay of many issues pertaining to the effectiveness of ITS. VanLehn (2011) reviewed randomized experiments that compared the effectiveness of human tutoring, computer tutoring, and no tutoring. He found that the effect size of ITS was 0.76, which was nearly as effective as human tutoring, $d = 0.79$. It appears that VanLehn (2011) found a larger effect than what the present meta-analysis revealed. However, we found that these two systematic reviews are different in at least three ways. First, the two reviews differ in subject domains and grade levels of students. The VanLehn (2011) review included studies of science, technology, engineering, and mathematics, with no restriction of grade levels. As a result, it included a large portion of studies on the use of ITS in college students' learning. The present meta-analysis focuses on the effectiveness of ITS on K–12 students' mathematical learning. Second, the two reviews had different methodological standards and applied different study inclusion criteria. VanLehn (2011) covered experiments that manipulated ITS interventions while controlling for other influences and excluded studies in which the experimental and comparison groups received different learning content. For example, it excluded all studies of Carnegie Learning's Cognitive Tutor because students in the experimental groups used a different textbook and classroom activities than did those in the comparison groups. In contrast, in the present meta-analysis, we placed no such restrictions. In fact, our meta-analysis includes studies that compared two ecologically valid conditions in which ITS may or may not be the only difference between the conditions. As a result, 20 out of the 31 independent studies included in this meta-analysis are studies of Cognitive Tutor. Last, VanLehn (2011) selected the outcome with the largest effect size in each primary study. The present meta-analysis extracted effect sizes for all the outcomes possible in each study and averaged them. Taken together, the differences mentioned above may help explain the seemingly discrepant findings from these two reviews. An overarching message from this is that when addressing the effectiveness of ITS, as is the case with many other educational interventions, one ought to ask a few questions: for whom, compared with what, and in what circumstances?

We compared the findings of the current meta-analysis with those of four recent reviews that focused on the effectiveness of computer technology or educational software on Pre-K to 12th graders' mathematical achievement. Methodologically, these reviews are also largely comparable with the current meta-analysis. In general, compared with the findings of some similar meta-analyses or systematic reviews of the effectiveness of educational technology, the effects of ITS appear to be relatively small.

Kulik (2003) reviewed 36 controlled evaluation studies to examine the effects of using instructional technology on mathematics and science learning in elementary and secondary schools. He found that the median effect of integrated learning systems was to increase mathematics test scores by 0.38 standard deviations, or from the 50th to the 65th percentiles. He also found that the median effect of computer tutorials was to raise student achievement scores by 0.59 standard deviations, or from the 50th to the 72nd percentiles.

Murphy et al. (2002) reviewed 13 studies of the efficacy of discrete educational software on Pre-K to 12th grade students' math achievement. They found that the overall weighted mean effect size for discrete educational software applications in math instruction was 0.45, and the median effect size was 0.27. On the basis of the distribution of confidence intervals, they concluded that $d = 0.30$ or greater appeared to be a reasonable estimate for the effectiveness of discrete educational software on mathematics achievement.

Slavin and Lake (2008) reviewed 38 studies to investigate the effects of CAI on elementary mathematics achievement. They found that the median effect size was 0.19. In their review of middle and high school math programs, Slavin, Lake, and Groff (2009) found that the weighted mean effect size was 0.10 for the effectiveness of CAI.

We should be quick to point out that conclusions based on the comparisons of the findings from different reviews ought to be tentative because there were variations among the reviews regarding the types of educational technology. As the use of educational technology became such a common practice in teaching and learning, it is increasingly difficult to picture a matrix of existing and ever-changing educational technology. As described in the introduction and Method section, we defined ITS as self-paced, learner-led, highly adaptive, and interactive learning environments operated through computers. In the studies included in this meta-analysis, ITS delivered learning content to students, tracked and adapted to students' learning paces, assessed learning progress, and gave students feedback. We believe these features distinguish ITS from other educational technologies in previous reviews.

One possible explanation for the small effects revealed in this meta-analysis is related to the degree of technology implementation and the purposes of technology use in classrooms. Evidence suggests that computer technology appears to have stronger effects when being used as supplemental tools than when used as the only or main instructions. For example, Schmid et al. (2009) found that in terms of degree of technology use, low ($g = 0.33$) and medium use of technology ($g = 0.29$) produced significantly higher effects than did high use ($g = 0.14$). They also found that in terms of type of technology use, when used as cognitive support (e.g., simulations), educational technology produced better results ($g = 0.40$) than when it was used as a presentational tool ($g = 0.10$) or for multiple uses ($g = 0.29$). Also, Tamim, Bernard, Borokhovski,

Abrami, and Schmid (2011) found that computer technology produced a slightly but significantly higher average effect size when used as supporting instruction than when it was used for direct instruction. Taken together, these findings imply that computer technology's major strengths may lie in supporting teaching and learning rather than substituting or replacing main instructional approaches or acting as a stand-alone tool for delivering content. However, further research is needed to reach a firm conclusion. Schmid et al. (2009) argued that future research ought to move away from the "yes-or-no" question and move to other issues, such as how much technology is desirable for improving student learning and how to best use technology to promote educational outcomes.

In addition, much research has supported the view that educational technology can improve student motivation and therefore positively influence student academic performance (Beeland, 2002; Roblyer & Doering, 2010; U.S. Department of Education, 1995). We speculate that as educational technology has become such a common part of learning environments in today's educational settings, student motivation and novelty effects related to the access of educational technology might have decreased. As a result, the relative effectiveness of educational technology may be declining.

Last, findings of this meta-analysis need to be interpreted with caution. As we described earlier, the results were largely based on the 31 independent studies that compared learning outcomes of instructions with an ITS component with those without one. This broad comparison covered four types of categorized situations that were different from one another to varying degrees. For studies in which ITS were the only difference between the treatment and comparison conditions, it is reasonable to conclude that the detected effectiveness difference can be attributed to ITS. However, when ITS were not the only difference between the conditions, differences in outcome measures cannot be attributed solely to ITS. For example, for studies of Cognitive Tutors, the treatment and comparison conditions could differ not only in the use of Cognitive Tutors but also in teachers or school environments. In such cases, it cannot be ruled out that the effectiveness of ITS is masked by the relative ineffectiveness of the other intervention components, such as teachers or school environments. It also could be the case that the effectiveness of the other intervention components is masked by the relative ineffectiveness of ITS. Taken together, this meta-analysis provides information regarding whether and how students' learning outcomes might differ depending on the involvement or absence of ITS from the instructions. However, one ought to be aware that the effectiveness differences may or may not be attributed solely to ITS.

Findings From Testing for Moderators

As summarized earlier, two robust findings stand out from this meta-analysis. The first finding was that the effects of ITS appeared to be greater when the intervention lasted for less than a school year than when it lasted for one school year or longer. We offer three possible explanations for this finding. First, it might be that the novelty factor wears off and students' motivation declines. This explanation would suggest that, just as is the case for many other interventions, too much of a good thing is not a good thing. Again, this brings us back to the important issue regarding how

much technology is desirable and how to best use technology to improving student learning. Second, when ITS were in regular or long-term use in schools, researchers usually had no or very little involvement in the actual use of ITS during the study. In other words, the degree of implementation might have impacted the effectiveness of ITS.

Third, some differences in the durations of interventions might be responsible for the differential effectiveness of ITS. Specifically, we found that a number of major characteristics of the studies in which ITS lasted for one school year or longer might account for the small effect sizes yielded in the studies. These studies, such as the studies of Cognitive Tutor (e.g., Campuzano et al., 2009; Dynarski et al., 2007), were more often based on big national samples; used more rigorous study methods, such as random assignment; and used more distal outcome measures, such as standardized achievement tests. In contrast, studies in which ITS lasted for short time or one semester generally produced bigger effect sizes for a number of reasons. For example, they studied relatively less known ITS, they were more often based on small sample sizes, they used less rigorous study methods, and they often used specifically designed or nonstandardized outcome measures. Many previous meta-analyses have concluded that the study differences mentioned above have an impact on the magnitude of effect sizes. Moderator analyses of this meta-analysis confirm this conclusion. We need to use caution in applying this finding to practices before further research is conducted and a firmer conclusion is reached.

The second finding was that ITS helped general students learn mathematical subjects more than it helped low achievers. One possible explanation is that ITS may function best when students have sufficient prior knowledge, self-regulation skills, learning motivation, and familiarity with computers. It is possible that general students have more of the characteristics needed to navigate ITS than low achievers do. Therefore, they benefited more from using ITS. For low achievers, classroom teachers, rather than ITS, might be better leaders, motivators, and regulators to help them learn. Research has found that there are differences in the ways that high achievers and low achievers used ITS and other computer-based instruction tools (Hativa, 1988; Hativa & Shorer, 1989; Wertheimer, 1990). For example, Hativa (1988) found that low achievers, more than high achievers, were prone to make software- and hardware-related errors when working with a CAI system. He further concluded that it was possible that high achievers were much more able than low achievers to adjust to a CAI learning environment so that they were able to benefit more from it.

This finding draws new attention to the debate regarding whether the use of computer technology actually widens the achievement gap between high achievers and low achievers, students of high and low learning aptitudes, students with advantaged and disadvantaged backgrounds, or White and minority students. The results from some longitudinal studies of CAI have provided support for the notion that computerized learning contributes to the increasing achievement gaps between students with different socioeconomic statuses, achievement levels, and aptitudes (Hativa, 1994; Hativa & Becker, 1994; Hativa & Shorer, 1989). Ceci and Papiero (2005) noted that nontargeted technology intervention that is used differently by advantaged and disadvantaged groups of students leads to achievement gap widening. This meta-analysis

adds further support to the above conclusions with the evidence that ITS might have contributed to the achievement gap between higher and lower achieving students. It is worth noting that only three studies provided results for low achievers.

This issue merits considerable attention. As mentioned earlier, the motivation of ITS development is to help students achieve learning gains as they do with the help of expert human tutors. There has been the expectation that ITS, as a form of advanced learning technology, ought to be able to provide optimal conditions needed to teach all children, given their interactivity, adaptability, and ability to provide immediate feedback and reinforcement. Developers of ITS may also want to consider way to adapt ITS for students with a variety of aptitudes and design culturally relevant technology learning environments. Further research with more nuanced approaches for ITS evaluation is needed to provide more information for this issue.

Conclusion

This meta-analysis synthesized studies of the relative effectiveness of ITS compared with regular class instruction on K–12 students' mathematical learning. Findings suggest that overall, ITS appeared to have no negative and perhaps a small positive impact on K–12 students' mathematical learning. The main contributions of this meta-analysis lie on three fronts. First, it provided further evidence for the conclusions that educational technology might be best used to support teaching and learning. Second, this meta-analysis revealed that ITS appeared to have a greater positive impact on general students than on low achievers. This finding will likely draw considerable attention in policy debates on the issue of whether computerized learning might contribute to the achievement gap between students with different achievement levels or prior backgrounds. Meanwhile, this finding implies that ITS research might be helpful in gaining a better understanding of how better learners learn through ITS, especially in terms of cognitive and metacognitive factors. Third, findings of this meta-analysis confirm several conclusions from many previous meta-analyses concerning the association between methodological features (e.g., sample size, research design, and outcome measure) of primary research and the effectiveness of the intervention studied. On the basis of the findings of this meta-analysis and similar reviews of educational technology, it seems best to think of ITS as one option in the array of education resources that educators and students can use to support teaching and learning. For students who are motivated and can self-regulate learning, ITS might be effective supplements to regular class instruction. However, ITS may not be efficient tools to boost low achievers' or at-risk students' achievement.

References

- References marked with an asterisk indicate studies included in the meta-analysis.
- Aleven, V., McLaren, B. M., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, 16, 101–128.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4, 167–207. doi:10.1207/s15327809jls0402_2
- *Arbuckle, W. J. (2005). *Conceptual understanding in a computer assisted Algebra I classroom* (Doctoral dissertation). Retrieved from ProQuest Information and Learning Company. (UMI No. 3203318)
- Arnott, E., Hastings, P., & Allbritton, D. (2008). Research Methods Tutor: Evaluation of a dialogue-based tutoring system in the classroom. *Behavior Research Methods*, 40, 694–698. doi:10.3758/BRM.40.3.694
- *Arroyo, I., Woolf, B. P., Royer, J. M., Tai, M., & English, S. (2010). Improving math learning through intelligent tutoring and basic skills training. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Lecture Notes in Computer Science: Vol. 6094. Intelligent tutoring systems* (pp. 423–432). Berlin, Germany: Springer-Verlag. doi:10.1007/978-3-642-13388-6_46
- *Beal, C. R., Arroyo, I. M., Cohen, P. R., & Woolf, B. P. (2010). Evaluation of AnimalWatch: An intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning*, 9, 64–77.
- *Beal, C. R., Waller, R., Arroyo, I., & Woolf, B. P. (2007). On-line tutoring for math achievement testing: A controlled evaluation. *Journal of Interactive Online Learning*, 6, 43–55.
- Becker, H. J. (1992). Computer-based integrated learning systems in the elementary and middle grades: A critical review and synthesis of evaluation reports. *Journal of Educational Computing Research*, 8, 1–41. doi:10.2190/23BC-ME1W-V37U-5TMJ
- Beeland, W. D., Jr. (2002). *Student engagement, visual learning and technology: Can interactive whiteboards help?* Retrieved from Valdosta State University website: http://chiron.valdosta.edu/are/Artmanscript/voll1nol/beeland_am.pdf
- *Biesinger, K., & Crippen, K. (2008). The impact of an online remediation site on performance related to high school mathematics proficiency. *Journal of Computers in Mathematics and Science Teaching*, 27, 5–17.
- Blok, H., Oostdam, R., Otter, M. E., & Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: A review. *Review of Educational Research*, 72, 101–130. doi:10.3102/00346543072001101
- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6), 4–16.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2006). *Comprehensive Meta-Analysis (Version 2.2.027)* [Computer software]. Englewood, NJ: Biostat.
- *Cabalo, J. V., Ma, B., & Jaciw, A. (2007). *Comparative effectiveness of Carnegie Learning's "Cognitive Tutor Bridge to Algebra" curriculum: A report of a randomized experiment in the Maui School District*. Palo Alto, CA: Empirical Education.
- *Cabalo, J. V., & Vu, M.-T. (2007). *Comparative effectiveness of Carnegie Learning's "Cognitive Tutor" Algebra I curriculum: A report of a randomized experiment in the Maui School District*. Palo Alto, CA: Empirical Education.
- *Campuzano, L., Dynarski, M., Agodini, R., & Rall, K. (2009). *Effectiveness of reading and mathematics software products: Findings from two student cohorts—Executive summary* (NCEE 2009-4042). Retrieved from U.S. Department of Education, Institute of Education Sciences, National Center for Education and Regional Assistance website: <http://ies.ed.gov/ncee/pubs/20094041/pdf/20094042.pdf>
- *Carnegie Learning. (2001a). *Cognitive Tutor research results: Freshman Academy, Canton City Schools, Canton, OH* (Cognitive Tutor Research Report OH-01-91). Pittsburgh, PA: Author.
- *Carnegie Learning. (2001b). *Cognitive Tutor results report: Freshman Academy, Canton City Schools, Canton, OH, 2001*. Retrieved from http://www.carnegielearning.com/static/web_docs/OH-01-01.pdf
- *Carnegie Learning. (2002). *Cognitive Tutor results report*. Pittsburgh, PA: Author.
- Ceci, S. J., & Papiero, P. B. (2005). The rhetoric and reality of gap closing: When the "have-nots" gain but the "haves" gain even more. *American Psychologist*, 60, 149–160. doi:10.1037/0003-066X.60.2.149

- Cheung, A., & Slavin, R. E. (2012). How features of educational technology programs affect student reading outcomes: A meta-analysis. *Educational Research Review*, 7(3), 198–215. doi:10.1016/j.bbr.2011.03.031
- Conati, C., & VanLehn, K. (2000). Further results from the evaluation of an intelligent computer tutor to coach self-explanation. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Lecture Notes in Computer Science: Vol. 1839. Intelligent tutoring systems* (pp. 304–313). Berlin, Germany: Springer-Verlag.
- Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.
- Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.
- Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. J. Gmysrasiewicz, & J. Vassileva (Eds.), *Lecture Notes in Artificial Intelligence: Vol. 2109. User modeling 2001* (pp. 137–147). Berlin, Germany: Springer-Verlag.
- Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin, Germany: Springer. doi:10.1007/978-3-642-58625-5
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56, 455–463. doi:10.1111/j.0006-341X.2000.00455.x
- *Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., . . . Sussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort—Report to Congress* (NCEE 2007-4005). Retrieved from U.S. Department of Education, Institute of Education Sciences website: <http://ies.ed.gov/ncee/pdf/20074005.pdf>
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Graesser, A. C., Conley, M., & Olney, A. (2011). Intelligent tutoring systems. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook: Vol. 3. Applications to learning and teaching* (pp. 451–473). Washington, DC: American Psychological Association.
- Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*, 21, 27–58.
- Haslam, M. B., White, R. N., & Klinge, A. (2006). *Improving student literacy: READ 180 in the Austin Independent School District, 2004–05*. Washington, DC: Policy Studies.
- Hativa, N. (1988). Computer-based drill and practice in arithmetic: Widening the gap between high- and low-achieving students. *American Educational Research Journal*, 25, 366–397. doi:10.3102/00028312025003366
- Hativa, N. (1994). What you design is not what you get (WYDINWYG): Cognitive, affective, and social impacts of learning with ILS—An integration of findings from six-years of qualitative and quantitative studies. *International Journal of Educational Research*, 21, 81–111. doi:10.1016/0883-0355(94)90025-6
- Hativa, N., & Becker, H. J. (1994). Integrated learning systems: Problems and potential benefits. *International Journal of Educational Research*, 21, 113–119. doi:10.1016/0883-0355(94)90026-4
- Hativa, N., & Shorer, D. (1989). Socioeconomic status, aptitude, and gender differences in CAI gains of arithmetic. *Journal of Educational Research*, 83, 11–21.
- *Hwang, G.-J., Tseng, J. C. R., & Hwang, G.-H. (2008). Diagnosing student learning problems based on historical assessment records. *Innovations in Education and Teaching International*, 45, 77–89. doi:10.1080/14703290701757476
- *Koedinger, K. R. (2002). Toward evidence for instructional design principles: Examples from Cognitive Tutor Math 6. In D. S. Mewborn, P. Sztajn, D. Y. White, H. G. Wiegel, R. L. Bryant, & K. Nooney (Eds.), *Proceedings of the Annual Meeting [of the] North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 21–49). Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.
- *Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30–43.
- Konstantopoulos, S., & Hedges, L. V. (2009). Analyzing effect sizes: Fixed-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 279–294). New York, NY: Russell Sage Foundation.
- Kulik, J. A. (2003). *Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say* (SRI Project No. P10446.001). Arlington, VA: SRI International.
- Lane, H. C., & VanLehn, K. (2005). Teaching the tacit knowledge of programming to novices with natural language tutoring. *Computer Science Education*, 15, 183–201. doi:10.1080/08993400500224286
- McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, 36, 222–233. doi:10.3758/BF03195567
- *Mendicino, M., Razaq, L., & Heffernan, N. T. (2009). A comparison of traditional homework to computer-supported homework. *Journal of Research on Technology in Education*, 41, 331–358.
- *Morgan, P., & Ritter, S. (2002). *An experimental study of the effect of Cognitive Tutor Algebra I on student knowledge and attitude*. Retrieved from the Carnegie Learning website: http://carnegielearning.com/web_docs/morgan_ritter_2002.pdf
- Murphy, R. F., Penuel, W. R., Means, B., Korbak, C., Whaley, A., & Allen, J. E. (2002). *E-DESK: A review of recent evidence on the effectiveness of discrete educational software*. Menlo Park, CA: SRI International.
- *Pane, J. F., McCaffrey, D. F., Slaughter, M. E., Steele, J. L., & Ikemoto, G. S. (2010). An experiment to evaluate the efficacy of Cognitive Tutor Geometry. *Journal of Research on Educational Effectiveness*, 3, 254–281. doi:10.1080/19345741003681189
- *Plano, G. S., Ramey, M., & Achilles, C. M. (2007, January). *Implications for student learning using a technology-based algebra program in a ninth-grade algebra course*. Paper presented at the 13th Annual Office of Superintendent of Public Instruction January Conference and High School Summit, Seattle, WA.
- *Radwan, Z. R. (1997). *Evaluation of the effectiveness of a computer assisted intelligent tutor system model developed to improve specific learning skills of special needs student* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 9729551)
- *Ritter, S., Kulikowich, J., Lei, P.-W., McGuire, C. L., & Morgan, P. (2007). What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. In T. Hirashima, U. Hoppe, & S. S.-C. Young (Eds.), *Frontiers in Artificial Intelligence and Applications: Vol. 162: Supporting learning flow through integrative technologies* (pp. 13–20). Amsterdam, the Netherlands: IOS Press.
- Roblyer, M., & Doering, A. (2010). *Integrating educational technology into teaching* (5th ed.). Boston, MA: Allyn & Bacon.
- Rowley, K., Carlson, P., & Miller, T. (1998). A cognitive technology to teach composition skills: Four studies with the R-Wise writing tutor. *Journal of Educational Computing Research*, 18, 259–296. doi:10.2190/KW4V-FJKD-L7J1-EFK0
- *Sarkis, H. (2004). *Cognitive Tutor Algebra I program evaluation: Miami-Dade County Public Schools*. Lighthouse Point, FL: Reliability Group.
- Schmid, R. F., Bernard, R. M., Borokhovski, E., Tamim, R., Abrami, P. C., Wade, C. A., . . . Lowerison, G. (2009). Technology's effect on achievement in higher education: A Stage I meta-analysis of classroom applications. *Journal of Computing in Higher Education*, 21, 95–109. doi:10.1007/s12528-009-9021-8

- *Shneyderman, A. (2001). *Evaluation of the Cognitive Tutor Algebra I program*. Unpublished manuscript, Office of Evaluation and Research, Miami-Dade County Public Schools, Miami, FL.
- Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1*, 51–77. doi:10.1080/1049482900010104
- Shute, V. J., & Zapata-Rivera, D. (2007). *Adaptive technologies* (Research Report RR-07-05). Princeton, NJ: Educational Testing Service.
- Slavin, R. E., & Lake, C. (2008). Effective programs in elementary math: A best-evidence synthesis. *Review of Educational Research, 78*, 427–515. doi:10.3102/0034654308317473
- Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research, 79*, 839–911. doi:10.3102/0034654308330968
- *Smith, J. E. (2001). *The effect of the Carnegie Algebra Tutor on student achievement and attitude in introductory high school algebra* (Unpublished doctoral dissertation). Virginia Polytechnic Institute and State University, Blacksburg, VA.
- *Stankov, S., Rosic, M., Zitko, B., & Grubisic, A. (2008). TEx-Sys model for building intelligent tutoring systems. *Computers & Education, 51*, 1017–1036. doi:10.1016/j.compedu.2007.10.002
- Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research, 81*, 4–28. doi:10.3102/0034654310393361
- U.S. Department of Education. (1995). Effects on students. In *Technology and education reform: Technical research report*. Retrieved from <http://www.ed.gov/pubs/SER/Technology/ch9.html>
- VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*, 227–265.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*, 197–221. doi:10.1080/00461520.2011.611369
- VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31*, 3–62. doi:10.1080/03640210709336984
- VanLehn, K., Jordan, P. W., Rosé, C. P., Bhembé, D., Böttner, M., Gaydos, A., . . . Srivastava, R. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), *Lecture Notes in Computer Science: Vol. 2363. Intelligent Tutoring Systems: 6th International Conference* (pp. 158–167). Berlin, Germany: Springer.
- *Wallis, R. L. (2005). *Effects of Web-based tutoring software on math test performance: A look at gender, math-fact retrieval ability, spatial ability and type of help* (Unpublished master's thesis). University of Massachusetts at Amherst, Amherst, MA.
- Wertheimer, R. (1990). The geometry proof tutor: An “intelligent” computer-based tutor in the classroom. *Mathematics Teacher, 83*, 308–317.
- What Works Clearinghouse. (2004, December). *What Works Clearinghouse topic report: Curriculum-based interventions for increasing K-12 math achievement—middle school*. Retrieved from Department of Education, Institute of Education Sciences, website: <http://www.eric.ed.gov/PDFS/ED485395.pdf>
- What Works Clearinghouse. (2007, May). *WWC intervention report middle school math: Cognitive Tutor Algebra I*. Retrieved from Department of Education, Institute of Education Sciences, website: http://www.aea9.k12.ia.us/documents/filelibrary/pdf/cognitive_tutor/WWC_Cognitive_Tutor_052907_3B8688D14AA44.pdf
- What Works Clearinghouse. (2008). *What Works Clearinghouse: Procedures and standards handbook* (Version 2.0). Retrieved from Department of Education, Institute of Education Sciences, website: http://ies.ed.gov/ncee/wWc/pdf/reference_resources/wwc_procedures_v2_standards_handbook.pdf
- What Works Clearinghouse. (2009, July). *WWC intervention report middle school math: Cognitive Tutor Algebra I*. Retrieved from Department of Education, Institute of Education Sciences, website: http://www.aea9.k12.ia.us/documents/filelibrary/pdf/cognitive_tutor/WWC_CogTutor_Report_July2009_B2A3C279D0481.pdf
- What Works Clearinghouse. (2010a, August). *WWC intervention report high school math: Carnegie Learning curricula and cognitive tutor software*. Retrieved from Department of Education, Institute of Education Sciences, website: http://ies.ed.gov/ncee/wWc/pdf/intervention_reports/wwc_cogtutor_083110.pdf
- What Works Clearinghouse. (2010b, March). *WWC intervention report middle school math: Plato Achieve Now*. Retrieved from Department of Education, Institute of Education Sciences, website: <http://ies.ed.gov/ncee/wwc/interventionreport.aspx?sid=378>
- Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6*, 413–429. doi:10.1037/1082-989X.6.4.413
- Woolf, B. P. (2009). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Burlington, MA: Kaufman.

Received November 10, 2011

Revision received February 18, 2013

Accepted March 4, 2013 ■

Using Student Interactions to Foster Rule–Diagram Mapping During Problem Solving in an Intelligent Tutoring System

Kirsten R. Butcher
University of Utah

Vincent Aleven
Carnegie Mellon University

In many domains, problem solving involves the application of general domain principles to specific problem representations. In 3 classroom studies with an intelligent tutoring system, we examined the impact of (learner-generated) interactions and (tutor-provided) visual cues designed to facilitate rule–diagram mapping (where students connect domain knowledge to problem diagrams), with the goal of promoting students' understanding of domain principles. Understanding was not supported when students failed to form a visual representation of rule–diagram mappings (Experiment 1); student interactions with diagrams promoted understanding of domain principles, but providing visual representations of rule–diagram mappings negated the benefits of interaction (Experiment 2). However, scaffolding student *generation* of rule–diagram mappings via diagram highlighting supported better understanding of domain rules that manifested at delayed testing, even when students already interacted with problem diagrams (Experiment 3). This work extends the literature on learning technologies, generative processing, and desirable difficulties by demonstrating the potential of visually based interaction techniques implemented during problem solving to have long-term impact on the type of knowledge that students develop during intelligent tutoring.

Keywords: problem solving, intelligent tutoring, diagrams, visual interaction

In domains such as geometry, chemistry, and physics, problem solving typically requires learners to move fluidly between problem-specific representations and domain-general principles (e.g., geometry theorems or physics principles) that govern problem-solving strategies and solutions. Early research on expertise has established that understanding how domain principles relate to problem-specific features is a key component of expert knowledge (Chi, Feltovich, & Glaser, 1981). However, novice students struggle to apply appropriate domain principles across a variety of individual problems and often are distracted by superficial aspects of problem representations (Lovett & Anderson, 1994; Ross, 1989). Even when provided with worked examples that demonstrate a step-by-step model of an expert solution, most students spontaneously engage only in superficial processing or passive examination of these examples (Atkinson & Renkl, 2007).

In geometry and other STEM (science, technology, engineering, and mathematics) domains, visual representations are a key aspect of specific problem situations. For example, in chemistry, visual representations (e.g., ball-and-stick diagrams, Lewis structure diagrams, symbolic structural formulae) are used to depict the position of atoms in a molecule as well as the bonds between them. In physics, diagrams often are used to depict the physical situation represented in a problem as well as the forces operating on the problem situation. In geometry, a diagram is used to represent the geometric context of a specific problem. Geometry diagrams depict the geometrical relationships of visual elements (e.g., lines, rays) that include the given information (e.g., known angles) necessary to solve a problem. In each of these domains, effective problem solving requires the learner to connect relevant domain principles to key aspects of the visual representation(s) depicting the specific, to-be-solved problem. Domain-level principles determine what aspects of the visual representation are relevant, and govern the problem-solving strategies that can be applied to the representation for a correct solution.

Let us consider an example from geometry. Domain-level principles (more specifically, geometry postulates and theorems) drive the calculation of to-be-solved numerical values based upon the geometric relationships provided in the problem diagram. For example, the linear pair postulate can be applied when two adjacent angles sit on a single line and share a common side. Figure 1 depicts a situation in which Angle 1 and Angle 2 are a linear pair. In a linear pair, if the measure of Angle 1 is known, Angle 2 can be solved by subtracting the measure of Angle 1 from 180° (since a line = 180°).

In this article, we use the term *rule* to refer to domain-level propositions that govern the selection and application of a correct step during problem solving. Thus, we define the process of

This article was published Online First September 9, 2013.

Kirsten R. Butcher, Department of Educational Psychology, University of Utah; Vincent Aleven, Human Computer Interaction Institute, Carnegie Mellon University.

This research was supported, in part, by funding provided by the Pittsburgh Science of Learning Center, which is funded by the National Science Foundation, Award SBE-0354420.

We wish to thank the following people: Octav Popescu, Carl Angiolillo, Grace Leonard, Michael Nugent, and Andy Tzou for their contributions to tutor development; Thomas Bolster for assistance in the development and scoring of experimental materials; and Colleen Conko and Mark Schoming for their assistance and support in the implementation of this research.

Correspondence concerning this article should be addressed to Kirsten R. Butcher, Department of Educational Psychology, University of Utah, 1705 Campus Center Drive, MBH 327, Salt Lake City, UT 84112. E-mail: kirsten.butcher@utah.edu

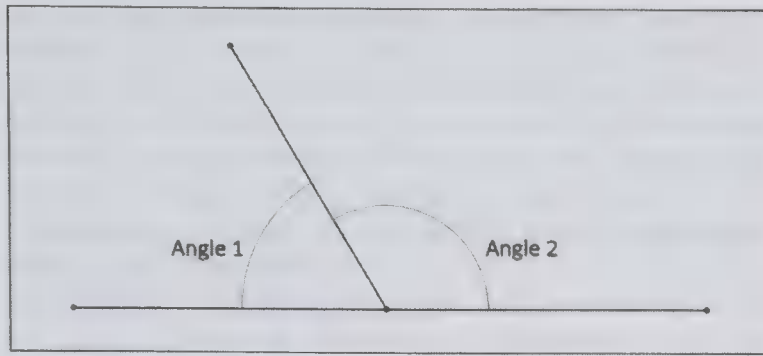


Figure 1. An example of a linear pair of angles depicted diagrammatically.

connecting domain-level rules (in this case, geometry postulates and theorems) to specific features of problem diagrams as *rule–diagram mapping*. Rule–diagram mapping represents a specific case of connecting domain-level principles to problem features. However, it should be noted that use of the term *rule* in this sense is not derived from the formal language of geometry; all geometry knowledge takes the form of definitions, postulates/axioms, or theorems. Postulates and axioms are statements about basic relationships or ideas in geometry that are accepted (or assumed) to be true (e.g., “Given any two points, there exists a line between them.”), and theorems are statements that have been proven to be true on the basis of definitions, postulates/axioms, or previously proven theorems (e.g., “Vertical angles are congruent.”). It is not clear the extent to which students understand or utilize these formal terms; although many instructional materials in geometry explain the terms *postulate* and *theorem*, some materials use informal language that may blur these distinctions. For example, some texts refer to the triangle inequality theorem as a rule (Carnegie Learning, 2007) or as a principle (Ryan, 2011). Across other texts, there are inconsistencies about what is named as a postulate or theorem. For example, a linear pair of angles is supplementary: Some texts call this the linear pair postulate (e.g., Carnegie Learning, 2010), and others call it the linear pair theorem (e.g., Carter, Cuevas, Day, Malloy, & Cummins, 2012). For simplicity’s sake, we refer to all domain-level reasons that students can use to justify geometry problem-solving steps as *rules*.

Strategies for Connecting Domain Principles and Specific Problem Features

Self-Explanation

Self-explanation is a robust learning strategy in which learners attempt to explain the content of learning materials to themselves, focusing on the meaning and importance of the instructional content as well as its connections to prior knowledge (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994; Renkl, Stark, Gruber, & Mandl, 1998). Chi et al. (1989) found that only about one quarter of students’ spontaneously produced self-explanations connected solution steps of physics problems to domain-level principles. Despite the relative rarity of their occurrence, the principle-based self-explanations that were generated served to enhance good solvers’ understanding of domain principles. Additional research has further distinguished between higher and lower quality self-explanations. Renkl (1997)

confirmed that—without specific prompting—few students go beyond passive or superficial explanation of examples. However, Renkl also noted that successful explainers tended to fall into two categories, one of which he termed *principle-based explainers*. These principle-based explainers focused on analyzing the subgoals of the example solution and elaborating on the principles related to those subgoals. Thus, these successful explainers connected domain principles to the specific steps of the example solution.

Although initial work on self-explanation focused on spontaneous processing, further research established that prompting self-explanation led to increases in understanding and more accurate mental models than in a control condition (Chi et al., 1994). Despite its potential support for learning, prompting the generation of high-quality self-explanations for large numbers of learners is a significant challenge. In computer environments, Hausmann and Chi (2002) found that typed self-explanations could be prompted as students worked with instructional materials on a computer, but these free-form, typed self-explanations were largely paraphrased statements that lacked quality. More promising results have been obtained with computer-supported interactions that structure the content of students’ self-explanations. Conati and VanLehn (2000) showed the effectiveness of user-adapted support for self-explanation in an intelligent tutoring system (ITS) for physics. Using prompts and drop-down menus, their system elicited self-explanations that connected problem-solving steps to domain principles and abstract solution plans. Using an ITS for geometry, Aleven and Koedinger (2002) showed that prompting students to name the high-level geometry principle (such as “corresponding angles”) that justified each problem-solving step improved the depth of student learning. Notably, when Aleven and Koedinger controlled for time on task (Experiment 2), benefits were limited to items requiring *understanding* of geometry rules (i.e., identifying unsolvable problems and naming geometry rules) rather than to the overall accuracy of numerical solutions (for which shallow problem-solving strategies can be more successful). In a system for learning probability, Atkinson, Renkl, and Merrill (2003) found support for near and far transfer when students were required to select (from a multiple-choice set of alternatives) the probability principle that justified each solution step (e.g., “multiplication principle”) in worked examples. Thus, scaffolding simple (verbally based) statements of problem-solving principles appears to be able to guide students toward deeper understanding of domain concepts.

Focusing Learner Processing on Central Problem Aspects

Naming the problem-solving principle associated with each step of a worked example or problem solution may be considered a fairly impoverished form of self-explanation compared to self-explanations uttered during natural, spontaneously generated speech. However, these self-explanation prompts may be effective because they focus student activity on key connections between the problem at hand and important domain concepts. Schworm and Renkl (2007) studied two types of self-explanation prompts in a system designed to train argumentation skills. In this research, the domain was “argumentation,” and domain-level principles (i.e., rules of argumentation) can be applied to specific content areas

(e.g., political topics). Schworm and Renkl (2007) developed self-explanation prompts that targeted either the specific content of the topic being argued (e.g., stem-cell research) or a domain-level principle of argumentation (e.g., provide an alternative theory). Only self-explanations related to the overarching domain principles promoted increased learning—self-explanations that focused on the specific content of the argument were ineffective. These findings are consistent with the general prescription that interactive features designed to support self-explanation in computer environments should target active processing of key conceptual or structural aspects of the to-be-learned domain (Atkinson & Renkl, 2007). This research also highlights the importance of thinking beyond specific problem instantiations to make connections to broader domain principles.

In domains where visual and verbal representations are central to problem solving, students' inability to understand the ways that domain principles relate to specific visual features of problems can result in misdirected attention and compromised learning. For example, Kozma (2003) found that chemistry students primarily focused on the surface features of representations and failed to connect visual features to underlying chemical principles. In contrast, chemistry experts utilized multiple representations (including diagrams, tables, and symbols) and made explicit connections between structural aspects of the representations and larger conceptual issues. Similarly, recent research has found that students perform well on chemistry questions that can be answered from visual information alone but poorly on questions that require the integration of information across representations such as a model and a graph (Stieff, Hegarty, & Deslongchamps, 2011). In physics, Wilkin (1997) found that problem-solving diagrams can decrease the effectiveness of self-explaining because students often use adjacency in diagrams to draw erroneous inferences that fail to be revised during learning. In geometry, novices also tend to focus on the surface-level similarities of diagrams. Lovett and Anderson (1994) found that students erroneously attempted to apply the same solution steps to geometry problems when diagrams looked similar but drew upon different logical structure. Lovett and Anderson concluded that in geometry—and other domains where diagrams are central to problem solving—the diagram serves as the basis for student recall. Thus, novice learners likely need support in understanding how key visual elements are tied to larger domain principles.

Despite students' apparent need for guidance in connecting visual representations to domain principles, self-explanations enacted in learning technologies—as discussed above—largely have taken the form of verbal statements that offer only weak support for visual-verbal connections between individual problems and domain principles.

Aleven and Koedinger (2002) noted that students who were prompted to name the geometry principle related to each problem-solving step in an ITS were more likely to perform well on hard-to-guess problems, which provided indirect evidence that they had developed more integrated visual-verbal knowledge during practice. However, these students showed much room for improvement. A key question is whether an ITS can improve student learning by helping students connect relevant visual and verbal elements. More specifically, can student learning in geometry be facilitated by an ITS that uses visually based interactive elements to connect (verbally expressed) domain principles and (visually represented) diagram features during problem solving?

Mapping Domain Principles to Problem Diagrams in Geometry

Ideally, self-explanations related to problem-solving diagrams should facilitate attention to or use of visual representations in ways that mimic expert processes, just as worked examples facilitate learning by prompting students to self-explain expert solution steps for a given problem (Atkinson, Derry, Renkl, & Wortham, 2000). Thus, it is important to consider how experts use geometry diagrams during problem solving. Koedinger and Anderson (1990) conducted research with experts solving geometry problems and, based upon findings from verbal data, developed a model of expert problem solving in geometry (the diagram configuration, or DC, model). Experts processed diagrams by identifying key configurations that were used to retrieve corresponding schematic knowledge. In the DC model, this was instantiated by parsing diagrams to identify key configurations (e.g., two parallel lines intersected by a transversal); the model used these configurations as well as given information about the problem to retrieve relevant schemas. Koedinger and Anderson's analysis of the DC model showed that it modeled expert processes well.

Getting students to recognize key configurations is particularly problematic in geometry, where problem-specific diagrams may vary widely in appearance even when the same key configurations are present. Moreover, superficial similarities in problem diagrams may mislead students into retrieving and applying similar geometry rules even when the problems contain a different underlying structure (Lovett & Anderson, 1994). Consider the examples in Figure 2. When two parallel lines are cut by a transversal, angles on the same side of the transversal and in the same relative position to the parallel lines are *corresponding angles* with congruent measures. Figure 2a shows two diagrams where the marked angles are corresponding. Figure 2b shows diagrams that are similar in

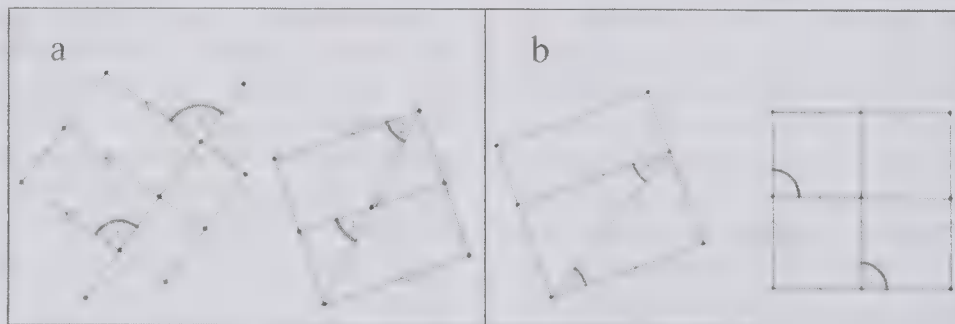


Figure 2. Diagrams with similar appearances: Angles 1 and 2 are *corresponding angles* in 2a, but not in 2b.

appearance, but the marked angles are not corresponding angles. If students understand the connection between the rule and the diagram in a vague or shallow way (i.e., in a square, angles that look similar are corresponding), they may incorrectly use *corresponding angles* as the (incorrect) justification to (correctly) answer that the angles in 2b are equal. Thus, they may get the numerical solution correct without understanding the underlying geometry rationale for their answers.

Developing connections between geometry rules and diagram configurations during problem solving is complicated by the fact that many geometry rules result in students using the same equation to generate a numerical solution, but for different logical reasons. For example, consider a problem where Angle X is given and the student's goal is to solve for Angle Y . If the angles are base angles of an isosceles triangle, Angle $X = \text{Angle } Y$. However, the same equation also will produce a correct answer if the angles are corresponding angles, alternate interior angles, or a pair of bisected angles. Thus, even in an ITS, students often develop shallow problem-solving strategies that result, behaviorally, in the observation that students are better at finding numerical answers to problems than explaining the reasons driving these answers (Aleven, Koedinger, Sinclair, & Snyder, 1998). The challenge is for students to move beyond these shallow (but often successful) strategies to develop an understanding of when and how to apply specific geometry rules to diagrams across a variety of problems. The purpose of this work is to explore interactive elements that support development of this rule-diagram mapping.

Facilitating Coordination of Visual-Verbal Information

A variety of methods have been used to try to facilitate connections between visual representations during problem solving and fundamental domain concepts or principles. Many of these methods can be grouped into two major categories: materials that physically coordinate visual and verbal information and materials that cue the learner to make connections between visual and verbal information sources.

Physical coordination. There is a great deal of research that has demonstrated that the ways in which visual and verbal information is combined have impacts on learning (e.g., Bodemer, Ploetzner, Feuerlein, & Spada, 2004; Glenberg & McDaniel, 1992; Mayer & Anderson, 1992; Moreno & Mayer, 1999; Tabbers, Martens, & van Merriënboer, 2004). Results have demonstrated that spatial contiguity (Mayer, 2001) between visual and verbal information reduces cognitive load by removing the effort associated with spatially mapping between information sources. However, it is possible to develop situations where requiring students to map between visual and verbal information sources can improve learning. Research has found that requiring students to actively integrate split source materials (e.g., by dragging text labels into a visual diagram) improves learning more than does providing learners with pre-integrated representations, especially when the materials are complex (Bodemer, Ploetzner, Bruchmüller, & Hacker, 2005). Further, recent research shows that learners who are provided with both concrete and abstract diagrams can transfer their knowledge better than learners provided with a single representation, a finding likely due to the implicit support that multiple representations provide for making connections between existing knowledge, current learning materials, and larger domain concepts

(Moreno, Ozogul, & Reisslein, 2011). Thus, instructional materials may support learning if interactions are used to facilitate mapping that connects relevant aspects of representations to larger domain concepts.

Visual cuing. Visual cues that focus learner attention on relevant features of representations during learning have been found to be quite powerful in supporting learning with multimedia materials. Eye-tracking evidence has shown that attending to important problem features can facilitate problem solving (Grant & Spivey, 2003), even when learners are not aware of their attentional focus (Thomas & Lleras, 2007). Multimedia research has found learning benefits when presentations direct learners' attention to relevant content by spotlighting visual features as they are discussed in an audio presentation (de Koning, Tabbers, Rikers, & Paas, 2007). In fact, de Koning and colleagues (de Koning, Tabbers, Rikers, & Paas, 2010) found no differences in learning resulting from learner-generated explanations and from provided instructional explanations when visual cues guided attentional focus during study. According to de Koning et al. (2010), visual cues may serve to increase active processing of instructional materials across a variety of explanation conditions. Since visual cues provide support in attending to central aspects of the learning materials, it is sensible to conclude that the combination of visual cues and explanation is effective because it concentrates processing on the most important aspects of the learning materials. However, it is an open question as to whether visual cues can, themselves, serve to facilitate effective processing and whether or not learner generation of such cues can enhance understanding.

The Current Research

Informed by the existing research outlined above, we address two key questions regarding computer-supported understanding of rule-diagram mapping. The first key question is: Can computer-supported interactions focused on key features of the visual diagram improve students' abilities to apply domain-general geometry "rules" to specific problem diagrams? If, as proposed by Lovett and Anderson (1994), diagrams serve as the basis for student recall, focusing student processing on key configurations of visual diagrams during problem solving may support understanding and long-term recall of rule-diagram connections. In this research, we explored two forms of visually targeted interactions: (a) self-explanations that were targeted to diagram features and (b) on-demand help that provided visual cues for rule-diagram mapping.

The second key question is: Who should generate the rule-diagram mappings, the tutor or the student? In an ITS (and in other forms of instruction), information that is central to the learning task can be either provided or withheld by the tutor (Koedinger & Aleven, 2007). For example, an ITS can highlight relevant information in a diagram, or it can require the student to highlight key visual features. Although overt activities that require the student to generate new information or representations should promote learning (Chi, 2009), a crucial factor is how successful the student will be in generating the targeted information without excessive floundering (Koedinger & Aleven, 2007). Thus, specific scaffolding may be required to structure the generation process and support students' development of meaningful representations. Another key consideration is how effective the generative activities will be in helping students process the connections between specific problem

features and domain-level principles (in this case, between diagram features and geometry rules).

We conducted a series of three experiments that explored the impact of different forms of rule–diagram mapping (verbal explanations or visual representations) as well as the instructional source of such mappings (student-generated or tutor-provided) on students' problem-solving success and learning of geometry rules. The first experiment focused on an explanation-based approach that required students to articulate specific problem features involved in the application of domain principles; students generated a written mapping of diagram features to geometry rules. The second experiment examined an alternate approach to identifying problem features involved in the application of domain principles; in this study, on-demand help provided a visual mapping between diagram features and geometry rules. The third experiment examined two different methods of visual mapping during problem solving: student-generated mappings versus tutor-provided mappings.

Experiment 1

In this experiment, we examined the effects of a relatively simple way of having students self-explain the rule–diagram mapping for each step during problem solving. As in Aleven and Koedinger (2002), all students identified the geometry rule that justified each problem-solving step during intelligent tutoring practice. However, a rule–diagram mapping factor was added to this self-explanation activity in which some students also went on to identify the specific diagram features that were relevant to the application of the named geometry rule. This experiment also varied the degree to which students' attention was focused on visual features in the geometry diagram during problem solving, by varying whether students interacted with problem diagrams or a solutions table during tutoring practice.

Method

Participants. Participants were 96 students from six 10th grade geometry classes at a vocational school in rural Pennsylvania. All classes were taught by the same teacher. Within each class, students were randomly assigned to the four experimental conditions.

Materials.

Geometry Cognitive Tutor. As a platform for our research, we used the Geometry Cognitive Tutor, one of several existing Cognitive Tutors (e.g., Aleven & Koedinger, 2002; Anderson, Corbett, Koedinger, & Pelletier, 1995). Cognitive Tutors are a type of ITS based on the ACT-R theory of cognition and learning (Anderson & Lebière, 1998); several studies have found that Cognitive Tutors are very effective in supporting student learning (Anderson et al., 1995; Koedinger, Anderson, Hadley, & Mark, 1997). The Cognitive Tutor uses algorithms and cognitive models to track students' skill development and to select practice problems for students. The Cognitive Tutor also uses a number of mechanisms to reduce cognitive load demands, including a step-by-step problem-solving sequence (where problem-solving subgoals are laid out for students) and immediate feedback at every step. None of these successful tutor features were manipulated in the current work.

The Geometry Cognitive Tutor is part of a full-year “hybrid” course in geometry that includes a text, ancillary materials, training for teachers, and the Cognitive Tutor software (Ritter, Anderson, Koedinger, & Corbett, 2007). Before participating in this research, all students had been using the Geometry Cognitive Tutor as part of their classroom curriculum for several months and were familiar with its basic functions (e.g., how feedback is displayed).

The research design for this experiment was a 2×2 factorial design that varied the locus of interaction during problem solving (an interactive diagram vs. a solutions table) and the self-explanation of rule–diagram mappings (no mapping vs. rule–diagram mapping). We first describe the two levels of the locus of interaction factor, followed by the two levels of the mapping factor.

Table interaction tutor. When the locus of interaction was the solutions table, all student interactions took place in a table separate from the geometry diagram (see Figure 3). Students entered answers and received tutor feedback in the table. As typical in the Geometry Cognitive Tutor, students needed to enter all values and rules correctly to complete a problem. Thus, students revised incorrect entries until correct. A static diagram (i.e., the diagram did not change in any way and students could not interact with it) was provided for each problem.

Diagram interaction tutor. When the locus of interaction was the diagram, students interacted directly with the geometry diagram as they worked in the Cognitive Tutor. The unknown (to-be-solved) quantities were represented in the diagram by question marks (see Figure 4). When students clicked a question mark, a small work area opened (co-located with the diagram) that allowed students to enter answers and receive feedback. As in the table interaction tutor, students needed to enter all values and rules correctly to complete a problem; students revised incorrect entries until correct. Correct numerical solutions were integrated directly into the diagram (see Figure 4).

No mapping. In this level of the mapping factor, the second experimental factor we varied, students entered only the numerical solutions and geometry rules for each problem-solving step (see Figure 5). Rules were either manually typed or selected from the tutor glossary.

Rule–diagram mapping. In the rule–diagram mapping condition, students were required to name the diagram features that were necessary to use the geometry rule that they had named for the problem-solving step. Necessary diagram features were defined as those that were used in the application of this rule. Because some postulates and theorems operate on multiple diagram features (e.g., the angle addition postulate requires the addition of two angles), the tutor scaffolded student answers by activating the number of “applied to” fields that corresponded to the number of arguments required in the selected rule. For example, in Figure 5, the central angle theorem requires only one argument (the known central angle or its intercepted arc), so only one “applied to” field is activated for completion. In order to expedite student identification of relevant diagram elements, students clicked on the quantities displayed in the diagram or the solutions table as a convenient (“one-click”) shorthand for naming the diagram features corresponding to these values (e.g., an angle or an arc). The tutoring software displayed the name of the corresponding diagram feature (e.g.,

The screenshot shows the Cognitive Tutor interface with a menu bar (File, Edit, Tutor, Go To, Window, Help) and a title bar (Cognitive Tutor). The main window is titled "Angles Table Reason Only Unit 1 Section 1 Arrowhead Demo". It features a "Look Back" button, a "Progress" indicator, and "Hint" and "Done" buttons.

DIAGRAM

A team of archaeologists on the Texas - Louisiana border excavated several broken arrowheads. Without the tell-tale feathered end of the arrow which has the tribal markings, the team couldn't decipher which tribe the arrows belonged to. Given the location, they know that they are either Choctaw or Cherokee arrows. To determine the tribal ancestry of the arrowheads, the archaeologists need to know how sharp of a point they have. History notes that the Choctaws were primarily an agricultural tribe, unaccustomed to making weapons aside from those to hunt, and therefore had arrowheads with wider points. The Cherokee, on the other hand, had many great warriors, and were skilled at making fine-pointed, fast arrows. (An angle sharper than 20 degrees is usually Cherokee.) However, the tip of the arrow was broken off and lost in transit. Help the archaeologists solve the mystery.

1 Dr. Sutton approximates that the corner of one of the arrows, angle ARO, equals 77 degrees. How sharp of a point does the arrowhead (m<WAR) have?

Hold button to see picture

REASONTOOL

	Value	Rule
m<ARO	77	Given
m<OWA	77	Isosceles Triangle
m<WAR	102	

When the locus of interaction is the solutions table, the diagram is visible but not interactive.

Answers and rules are entered in a solutions table, where tutor feedback also is given. Correct answers are displayed in the table but do not appear in the diagram.

Figure 3. An annotated screenshot of the Cognitive Tutor where the table is the locus of interaction.

Arc EO) in a dedicated field in the open work area (or the solutions table) called the "Applied To" field.

For example, as can be seen in Figure 5, the student has named the "Central Angle" theorem as the principle used to find the measure of Angle OTE. Since a central angle is equal to the measure of its intercepted arc, the student needs to indicate the intercepted arc. To do so, the student clicks the solved value of 85.8 for the arc (either in the diagram or the table, depending upon condition), which enters "Arc EO" in the "Applied To" field. Students received immediate feedback on named diagram elements, as for all other submitted answers in the Cognitive Tutor. As with all Cognitive Tutor answers, students were required to revise any incorrect answers until all entries were completed correctly.

Assessments: Pretest, immediate posttest, and delayed posttest. Student learning was assessed via three assessments: a pretest, a posttest, and a delayed posttest. The pretest and immediate posttest consisted of the same problems but in different orders to minimize superficial recognition; the delayed posttest was composed of new problems. The pre- and posttest contained 16 problems (two problems for each of eight geometry diagrams). The delayed posttest contained eight problems (two problems for each of four geometry diagrams). Although the posttest and delayed posttest targeted geometry rules from the same unit of

study, the delayed posttest included somewhat less complex diagrams with fewer embedded shapes (see Figure 6). All tests were administered individually by computer, using the tutoring software but without any tutoring (i.e., no feedback or hints); answers were recorded in software logs.

Solvability decisions. For each problem, students first were required to make a *solvability decision*, that is, to determine if a learned geometry principle would allow them to solve the problem with the available information. This solvability decision required students to reason carefully about the diagram features needed to apply a geometry rule and to determine if one of these rules was relevant to the existing problem. Solvability judgments are challenging in that they require students to consider all potentially relevant geometry rules and diagram relationships before answering "no." Students received 1 point for each correct solvability decision (pre/posttest maximum = 16, delayed posttest maximum = 8).

Numerical solutions. These items tested students' abilities to generate the correct numerical solution for to-be-solved angles (e.g., Angle ABC = 60°). Students received 1 point for every correctly solved item. (Due to the inclusion of solvability decisions, not all items had numerical solutions: pre/posttest maximum = 12; delayed posttest maximum = 5).

Diagram Text:

A team of archaeologists on the Texas - Louisiana border excavated several broken arrowheads. Without the tell-tale feathered end of the arrow which has the tribal markings, the team couldn't decipher which tribe the arrows belonged to. Given the location, they know that they are either Choctaw or Cherokee arrows. To determine the tribal ancestry of the arrowheads, the archaeologists need to know how sharp of a point they have. History notes that the Choctaws were primarily an agricultural tribe, unaccustomed to making weapons aside from those to hunt, and therefore had arrowheads with wider points. The Cherokee, on the other hand, had many great warriors, and were skilled at making fine-pointed, fast arrows. (An angle sharper than 20 degrees is usually Cherokee.) However, the tip of the arrow was broken off and lost in transit. Help the archaeologists solve the mystery.

Problem Statement:

1. Dr. Sutton approximates that the corner of one of the arrows, angle ARO, equals 77 degrees. How sharp a point does the arrowhead ($m\angle WAR$) have?

Diagram: A triangle with vertices A, R, and O. Angle ARO is labeled 77. Angle WAR is the angle to be solved.

REASONTOL Table:

Value	Rule
$m\angle ARO$	77
$m\angle OWA$	
$m\angle WAR$	103

Annotations:

- Correct numerical solutions are displayed in the diagram.
- All interaction takes place in or near the diagram. Students click a question mark to open a small, moveable work area near the to-be-solved item. Answers and rules are entered in the work areas, where tutor feedback also is given.
- When the locus of interaction is the diagram, the table is non-interactive. The table shows a passive record of answers and geometry rules.

Figure 4. An annotated screenshot of the Cognitive Tutor where the diagram is the locus of interaction.

Rule application. For each problem, students were asked to name the geometry rule that they used to derive their numerical solution and to map the features of the geometry diagram to the geometry rule that they had named (e.g., corresponding angles: Angle ABC = Angle FGD). Students received 1 point for each correctly identified rule and for each correct mapping to diagram features (pre/posttest maximum = 32; delayed posttest maximum = 16).

Statistical analyses. For each of the three experiments reported here, a series of three analyses of variance (ANOVAs) was conducted to assess the impact of experimental interventions on different types of student knowledge. A separate analysis was conducted for each of the three major types of assessment items: solvability decisions, numerical solutions, and rule application. In Experiment 1, a series of three repeated-measures ANOVAs were conducted where the independent variables were locus of interaction (table vs. diagram) and mapping (no mapping vs. rule–diagram mapping) and the repeated factor was test time (immediate posttest, delayed posttest). A Bonferroni correction was used to adjust alpha levels for multiple comparisons; analyses of student outcomes used an alpha level of $p = (.05/3) = .017$.

Because we were most interested in whether students were able to apply geometry rules accurately during problem solving, we analyzed student outcomes based on the percent correct of attempted answers (much like tests of cognitive skills that assess performance based upon accuracy of attempted items; cf. Hegarty & Waller, 2004). Students can skip problems for many reasons—they may have run out of time, missed an item as they worked, or determined that they didn't know the answer. By analyzing percent correct of attempted items, we assessed how well students were able to apply their geometry knowledge when we knew that they attempted to do so. Percent correct of attempted responses was calculated by dividing the number of correct answers by the total number of attempted answers; tables of means and standard deviations also report raw performance rates (percent correct) and attempt rates (percent attempted) for each condition across the assessment types (see Table 1).

Procedure. During the experiments described here, students participated in the study as part of their normal classroom activities, using the Geometry Cognitive Tutor for one (75 min) classroom block each week. In the first week of the study, students spent up to 30 min completing the pretest. Students spent three

No Mapping:
Students name only the geometry rule used in the problem-solving step.

Rule-Diagram Mapping:
For each problem-solving step, students must name not only the geometry rule but also the diagram features used in the application of the rule.

DIAGRAM
In circle T shown here, the measure of arc EO is equal to 85.8 degrees

REASONTOOL

	Value	Rule	Applied to	Applied to	Applied to
Arc EO	85.8	Given			
$m\angle EMO$					
$m\angle OTE$	85.8	Central Angle			
$m\angle MTG$					
Arc MG					
$m\angle MOG$					
$m\angle GTE$					
$m\angle OTM$					

Find:
Central Angle
Chord Product
Circle Area
Circle Fraction
Circumference
Congruent Chords
Congruent Radii

Interior Angle
Definition:
The measure of an interior angle in a circle is equal to half of the sum of the measures of the interior angle's intercepted arc and the intercepted arc of the interior angle's vertical angle.
Example:
Line m and line n intersect at the point E inside the circle M. Angle AEB (angle 1) is an interior angle of circle M. arc AB is the intercepted arc of the interior angle. arc CD is the intercepted arc of the interior angle's vertical angle.
measure of arc AB + measure of arc CD
2

Figure 5. An annotated screenshot of the Cognitive Tutor's rule-diagram mapping conditions.

classroom blocks (one per week in Weeks 2–4 of the study) working in the angles units of the Geometry Cognitive Tutor using their randomly assigned condition. In Week 5, students took the immediate posttest (30 min). One month following the posttest, students completed the delayed posttest (20 min).

Results

We limited our analyses to the 53 students who completed both the posttest and the delayed posttest. Sample size was comparable across conditions, as can be seen in Table 1, which shows the

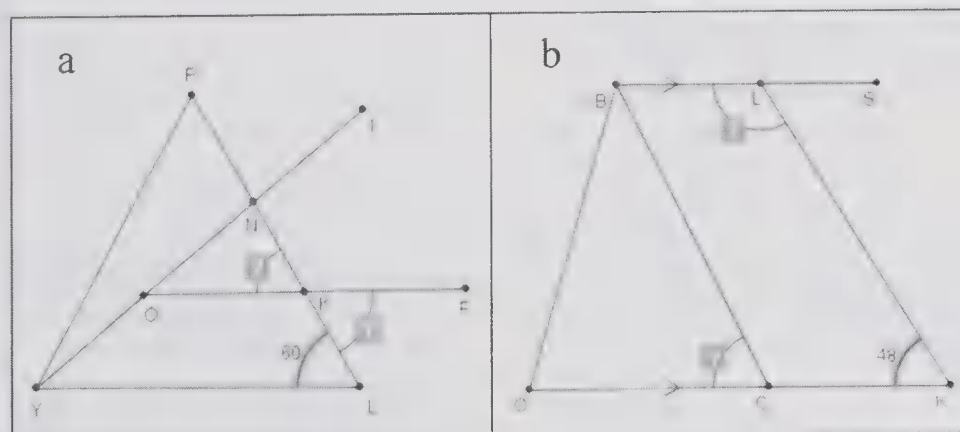


Figure 6. Example diagrams from a posttest item (6a) and a delayed posttest item (6b).

Table 1
Experiment 1 Means and (Standard Deviations) for Posttest and Delayed Posttest Assessment Items

Item type	Table interaction		Diagram interaction	
	No mapping	Mapping	No mapping	Mapping
Cognitive Tutor position prior to study	18.2 (3.6)	19.2 (3.5)	18.1 (5.2)	17.8 (4.8)
Posttest				
Solvability decisions	(<i>n</i> = 10)	(<i>n</i> = 14)	(<i>n</i> = 15)	(<i>n</i> = 14)
% correct	51.3 (15.8)	57.1 (19.3)	63.0 (12.6)	61.3 (18.3)
% attempted	96.3 (11.9)	96.4 (11.7)	98.2 (6.7)	100 (0)
% correct of attempted	53.9 (16.6)	59.5 (18.5)	64.1 (12.1)	61.3 (18.3)
Numerical solutions ^a	(<i>n</i> = 9)	(<i>n</i> = 12)	(<i>n</i> = 15)	(<i>n</i> = 14)
% correct	31.7 (22.2)	33.3 (34.9)	41.1 (32.6)	38.9 (30.0)
% attempted	46.3 (17.2)	54.5 (27.7)	68.8 (24.4)	57.9 (26.5)
% correct of attempted	47.6 (29.8)	38.3 (34.1)	48.0 (30.0)	51.9 (33.7)
Rule application	(<i>n</i> = 10)	(<i>n</i> = 14)	(<i>n</i> = 15)	(<i>n</i> = 14)
% correct	19.4 (15.5)	30.4 (22.3)	34.6 (18.6)	33.1 (23.5)
% attempted	96.3 (11.9)	96.2 (12.5)	98.0 (7.5)	100 (0)
% correct of attempted	20.2 (15.6)	31.0 (21.6)	34.9 (18.1)	32.9 (23.3)
Delayed posttest				
Solvability decisions	(<i>n</i> = 10)	(<i>n</i> = 14)	(<i>n</i> = 15)	(<i>n</i> = 14)
% correct	51.3 (19.9)	42.0 (19.4)	51.8 (23.4)	57.5 (16.2)
% attempted	95.8 (14.1)	96.1 (21.7)	96.4 (13.4)	100 (0)
% correct of attempted	53.8 (17.7)	48.4 (23.4)	53.6 (22.2)	57.5 (16.2)
Numerical solutions ^a	(<i>n</i> = 9)	(<i>n</i> = 12)	(<i>n</i> = 15)	(<i>n</i> = 14)
% correct	24.0 (28.0)	25.7 (31.8)	31.4 (28.0)	34.7 (29.7)
% attempted	47.5 (26.9)	45.5 (30.9)	57.1 (25.8)	58.3 (27.0)
% correct of attempted	34.7 (36.2)	39.5 (39.4)	38.0 (33.7)	44.4 (36.7)
Rule application	(<i>n</i> = 10)	(<i>n</i> = 14)	(<i>n</i> = 15)	(<i>n</i> = 14)
% correct	23.8 (19.5)	23.2 (14.2)	21.4 (17.3)	28.3 (25.8)
% attempted	95.0 (15.8)	92.0 (21.7)	96.4 (13.4)	100 (0)
% correct of attempted	24.2 (19.0)	25.7 (15.4)	21.0 (16.7)	28.2 (25.8)

^a One student in the table interaction/no mapping condition and two students in the table interaction/mapping condition attempted no numerical solutions and were dropped from analyses using percent correct of attempted.

means and standard deviations for assessment data. Although high, this rate of attrition is consistent with other studies conducted at the school that have experienced over 60% attrition (Salden, Aleven, Schwonke, & Renkl, 2010; Walker, Rummel, & Koedinger, 2009).¹

Due to a server error, pretest responses were not saved for 31 students and pretest data therefore were not used in analyses. As a check of random assignment, we conducted a two-way ANOVA where the number of units completed in the Cognitive Tutor prior to the start of the current study was the dependent variable and the locus of interaction and mapping factors were independent variables. Results showed no significant main effects or interactions ($F_s < 1$). Since students' classroom grades were calculated largely based upon their progress in the tutor, these results suggest that initial classroom performance was equivalent across conditions.

Learning outcomes.

Solvability decisions. Results showed no significant main effect of test time ($F_{(1, 49)} = 3.5, p = .07, \eta_p^2 = .07$) and no main effects of locus of interaction ($F_{(1, 49)} = 1.9, p = .18, \eta_p^2 = .04$) or mapping ($F < 1$). There were no significant two-way interactions ($F_s < 1$). The three-way interaction among test time, locus of interaction, and mapping was not significant ($F_{(1, 49)} = 1.7, p = .20, \eta_p^2 = .03$).

Numerical solutions. Results showed no significant main effect of test time ($F_{(1, 46)} = 1.9, p = .17, \eta_p^2 = .04$) and no main

effects of locus of interaction or mapping ($F_s < 1$). There were no significant two-way interactions ($F_s < 1$). The three-way interaction among test time, locus of interaction, and mapping also was not significant ($F < 1$).

Rule application. Results showed no significant main effect of test time ($F_{(1, 49)} = 2.3, p = .14, \eta_p^2 = .05$) and no significant main effects of locus of interaction or mapping ($F_s < 1$). There were no significant two-way interactions: test time by locus of interaction ($F_{(1, 49)} = 1.7, p = .20, \eta_p^2 = .03$), test time by mapping ($F < 1$), locus of interaction by mapping ($F < 1$). The three-way interaction was not significant ($F_{(1, 49)} = 2.0, p = .17, \eta_p^2 = .04$).

Discussion

Overall, the results from Experiment 1 showed that a simple form of self-explanation targeted to rule–diagram mapping—that

¹ In this study, the relatively high absentee rate likely is due to at least two factors. First, although students completed trade classes and the Pennsylvania state mathematics core at the vocational school, all other courses were completed at a traditional high school in the student's home school district. Thus, individual school schedules and special activities contributed to absences at the vocational school. Second, the timing of the study was determined by the course schedule of curriculum topics and resulted in the delayed posttest being given the first week following winter vacation—a likely contributor to high rates of student absence.

is, clicking a numerical quantity to “name” the diagram features to which a geometry rule applied—did not support understanding of these rules. Further, the locus of interaction did not demonstrate a significant impact on student learning.

Unlike in previous research (Butcher & Aleven, 2007), we did not see a significant benefit for diagram interaction when considering student outcomes. However, Butcher and Aleven (2007) used only a posttest and analyzed diagram mappings separately from rule naming. A post hoc analysis of the current data at posttest showed results that were weak but generally consistent with this prior work. We extracted students’ posttest *diagram mapping* scores from the rule application variable and analyzed these data using a two-way ANOVA where locus of interaction and mapping were the independent variables. Results showed nonsignificant trends for locus of interaction ($F_{(1, 49)} = 3.6, p = .06, \eta_p^2 = .07$) and for the interaction between locus of interaction and mapping ($F_{(1, 49)} = 3.4, p = .07, \eta_p^2 = .07$); there was not a significant main effect of mapping ($F < 1$). Similar to the rule application scores in Table 1 were findings that students who interacted with diagrams were best able to name relevant diagram features when they did not indicate mappings (diagram interaction with no mapping > diagram interaction with mapping), whereas the opposite was true for students who interacted with solutions tables (table interaction with no mapping < table interaction with mapping). Although these post hoc results are not strongly conclusive, it is possible that “stating” the mappings distracted students from their principal learning goal of understanding diagram configurations as related to geometry rules, much like requiring learners to respond to example gaps in addition to self-explanation prompts has been found to be detrimental to learning (Hilbert, Renkl, Kessler, & Reiss, 2008).

Why did this particular implementation of rule–diagram mapping fail to show benefits? One possibility may be that the rule–diagram mappings that the tutor elicited were redundant with processing that occurred as students determined the numerical solution and geometry rule for each problem-solving step. Rule–diagram mappings required students to indicate relevant diagram features by clicking on solved quantities displayed in the diagram, but students already attended to these quantities while generating the correct numerical solution. This explanation is supported by the fact that during training, rule–diagram mappings were correct 88% of the time.

Another possibility is that the mapping implemented in this study was too closely tied to specific aspects of individual problems (e.g., specific angle names) as opposed to general diagram configurations (e.g., two parallel lines intersected by a transversal). As noted elsewhere (Atkinson & Renkl, 2007), effective prompts must direct students’ attention to domain-level representations (Schworm & Renkl, 2007) rather than the specific content of individual problems. What would constitute domain-general mappings in geometry? Since geometry postulates and theorems are tied to key visual configurations in problem diagrams, visually representing these configurations (rather than naming specific diagram features) may be a better method of prompting rule–diagram mapping. Thus, we explored an alternative, visually based approach to mapping between geometry rules and diagram features in Experiment 2. In this study, on-demand help provided a visual mapping between rules and diagrams for students.

Experiment 2

In this experiment, we examined the impact of rule–diagram mapping implemented as visual cues (in the form of diagram highlights) within the on-demand help system of the Cognitive Tutor. These highlights mapped visual features of the geometry diagram to verbal references in the text-based hints. Unlike Experiment 1, where explanations focused on specific features (e.g., Angle ABC) of a problem representation, the visual highlighting in Experiment 2 cued the key diagram configurations (e.g., parallel lines) relevant to geometry rules across a variety of problem representations.

Method

Participants. Participants were 109 students from seven 10th grade geometry classes at the same vocational school as in Experiment 1. All classes were taught by the same teacher; both the teacher and the students were different from those in Experiment 1. Students in each class were randomly assigned to one of the four experimental conditions described below.

Materials.

Geometry Cognitive Tutor. As in Experiment 1, the Geometry Cognitive Tutor was used to vary the locus of interaction (diagram vs. table). In Experiment 2 we also varied the visual appearance of the on-demand hints (highlighted vs. not highlighted).

Hint format: Highlighted hints versus standard hints. The purpose of the highlighted hints was to provide students with a visual mapping between diagram features and the geometry rules used during problem solving. The highlighted diagram features are key to the rule’s applicability conditions. These rule–diagram mappings were implemented via step-by-step explanations provided in the on-demand hints. Multiple hint levels are available for every subgoal in the Cognitive Tutor at the student’s request. The highlighted hints provided learners with a color-coded visual mapping between the text referents present in the explanation of the geometry rule and the geometry diagram for the current problem; the color-coded highlighting was updated as the hint text changed when learners continued through the hints (see Figure 7).

Standard hints provided students with the same explanations (i.e., identical text) in the same order as the highlighted condition. However, neither the standard hints nor the accompanying problem diagrams were highlighted to show connections between geometry rules and the specific diagrams. Standard hints appeared as plain text.

Assessments. Assessments were similar to those used in Experiment 1 except that a delayed posttest wasn’t possible because the circles units targeted by Experiment 2 were positioned at the end of the academic school year. There was not enough time left in the academic calendar for a delayed posttest to be implemented following the posttest.

Small changes also were made in the assessments in response to the teacher’s preference that assessments be given on paper rather than on the computer. Whereas the computer interface used in Experiment 1 locked irrelevant answer areas once the student made a solvability decision, pilot testing showed that students using the paper test typically attempted to complete all blanks. Thus, solvability decisions in this experiment were implemented as a series of true/false statements (e.g., “You can use the *inscribed angle* rule to find the measure of arc KOP if you know only the

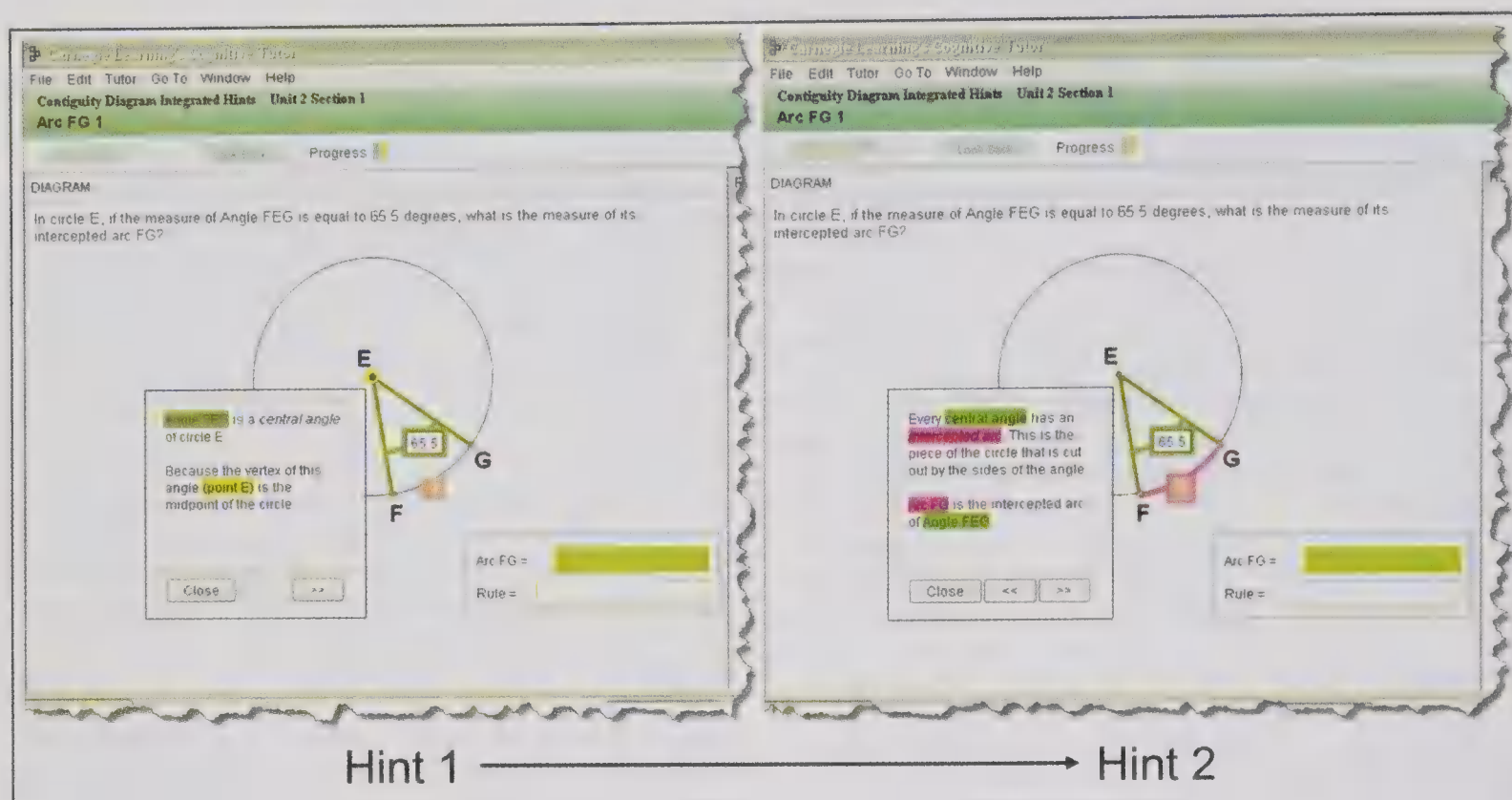


Figure 7. Highlighting in two hints for a problem-solving step (find the measure of Arc FG).

measure of angle KNP.”). Three diagrams were presented with six solvability items per diagram; students received 1 point per correct solvability decision (maximum = 18 points).

Numerical solutions and rule application were tested by 15 geometry problems that made use of three problem diagrams. As in Experiment 1, students received 1 point per correct response (numerical solutions maximum = 15; rule application maximum = 30).

Statistical analyses. Since there was no delayed posttest in Experiment 2, learning outcomes were analyzed using a series of two-way ANOVAs (one for each assessment category: solvability decisions, numerical solutions, and rule application). As in Experiment 1, a Bonferroni correction was used to correct for multiple comparisons. Alpha levels were set at $p = (.05/3) = .017$.

Procedure. The procedure was the same as in Experiment 1 except that the study ended with a paper posttest and did not include the delayed posttest.

Results

Of the 109 students enrolled in the seven geometry classes, 101 students took the pretest. Of these students, four failed to attempt any solvability decisions, 11 failed to attempt any numerical solutions, and 19 failed to attempt any rule application items. As a check of random assignment, three two-way ANOVAs (where locus of interaction and hint format were independent variables) were used to examine pretest performance on solvability decisions, numerical solutions, and rule application. Means and standard deviations for pretest scores are shown in Table 2. Results showed no significant main effects of experimental conditions on assessment item types: pretest solvability decisions (locus of interaction: $F_{(1, 93)} = 1.4, p > .24$; hint format: $F_{(1, 93)} = 1.5, p > .25$; interaction: $F < 1$), pretest numerical solutions (locus of interaction, hint format, and interaction: $F_s < 1$), and pretest

rule application (locus of interaction: $F_{(1, 78)} = 2.0, p > .16$; hint format: $F_{(1, 78)} = 2.0, p > .16$; interaction: $F < 1$). Thus, pretest data are not considered further.

Of the 101 students who took the pretest, 77 also took the posttest. Of these students, two students failed to attempt any numerical solutions (leaving 75 for analysis), and 10 failed to attempt any rule application items (leaving 67 for analysis). No student failed to attempt solvability decisions, leaving all 77 students for analysis. Attrition was comparable across conditions (see Table 2). Table 2 shows the means and standard deviations for assessment items by experimental condition.

Solvability decisions. Results showed no significant main effect of locus of interaction ($F < 1$) or hint format ($F_{(1, 73)} = 3.7, p = .06, \eta_p^2 = .05$). The interaction between locus of interaction and hint format was not significant ($F < 1$).

Numerical solutions. Results showed no significant main effect of locus of interaction or hint format ($F_s < 1$). The interaction between locus of interaction and hint format was not significant ($F_{(1, 71)} = 1.2, p = .28, \eta_p^2 = .02$).

Rule application. Although results showed no significant main effect of locus of interaction ($F_{(1, 63)} = 3.5, p = .07, \eta_p^2 = .05$) or hint format ($F < 1$), the interaction between locus of interaction and hint format was statistically significant ($F_{(1, 63)} = 6.4, p = .014, \eta_p^2 = .09$; see Table 2). Follow-up analyses (F calculated with the error term from the interaction and applying Bonferroni correction) showed that there was a significant difference between the locus of interaction conditions when viewing the standard hints ($F_{(1, 63)} = 9.29, p = .003$), but there was not a significant difference between locus of interaction conditions when viewing the highlighted hints ($F < 1$). As can be seen in Figure 8, students who saw the standard hints were most successful

Table 2

Experiment 2 Means (and Standard Deviations) for Pre- and Posttest Assessment Items

Item type	Table interaction		Diagram interaction	
	Standard hints	Highlighted hints	Standard hints	Highlighted hints
Pretest				
Solvability decisions (true/false items)	(<i>n</i> = 24)	(<i>n</i> = 25)	(<i>n</i> = 24)	(<i>n</i> = 24)
% correct	56.0 (57.0)	43.3 (17.3)	46.5 (17.9)	46.8 (13.2)
% attempted	80.8 (28.8)	79.8 (25.4)	84.5 (25.6)	86.3 (20.4)
% correct of attempted	68.1 (53.5)	55.1 (13.2)	55.3 (14.9)	53.9 (25.3)
Numerical solutions	(<i>n</i> = 20)	(<i>n</i> = 25)	(<i>n</i> = 24)	(<i>n</i> = 21)
% correct	16.3 (14.7)	17.3 (15.4)	15.8 (15.1)	15.9 (12.7)
% attempted	48.7 (32.7)	43.2 (29.0)	44.4 (31.6)	44.1 (30.8)
% correct of attempted	38.2 (29.4)	40.0 (28.3)	40.4 (27.3)	44.1 (34.2)
Application of principles	(<i>n</i> = 18)	(<i>n</i> = 23)	(<i>n</i> = 22)	(<i>n</i> = 19)
% correct	3.7 (5.5)	3.8 (5.0)	6.2 (6.8)	3.7 (4.0)
% attempted	25.7 (23.8)	28.6 (21.6)	26.7 (22.6)	33.2 (28.3)
% correct of attempted	21.0 (27.3)	12.3 (16.9)	28.5 (28.9)	21.0 (30.3)
Posttest				
Solvability decisions (true/false items)	(<i>n</i> = 22)	(<i>n</i> = 20)	(<i>n</i> = 15)	(<i>n</i> = 20)
% correct	51.0 (11.8)	50.6 (12.3)	47.8 (19.9)	51.4 (35.3)
% attempted	88.1 (22.3)	92.8 (15.1)	79.3 (29.1)	92.8 (22.3)
% correct of attempted	61.5 (18.7)	55.0 (11.6)	62.5 (21.8)	55.0 (9.3)
Numerical solutions	(<i>n</i> = 18)	(<i>n</i> = 22)	(<i>n</i> = 15)	(<i>n</i> = 20)
% correct	24.9 (16.6)	33.7 (45.3)	32.0 (61.8)	26.7 (21.3)
% attempted	62.4 (26.3)	70.0 (31.5)	64.4 (35.2)	73.3 (34.2)
% correct of attempted	45.8 (32.3)	56.7 (29.5)	49.9 (28.6)	45.0 (32.6)
Application of principles	(<i>n</i> = 21)	(<i>n</i> = 15)	(<i>n</i> = 12)	(<i>n</i> = 19)
% correct	6.0 (8.3)	11.8 (9.1)	16.7 (13.6)	9.1 (7.6)
% attempted	26.4 (17.5)	47.3 (31.0)	42.2 (32.7)	44.7 (30.6)
% correct of attempted	15.9 (20.7)	31.2 (27.8)	43.3 (28.8)	27.1 (24.3)

in applying geometry rules to diagrams at posttest when they had interacted directly with the diagrams during intelligent tutoring practice. When students were provided with rule–diagram mappings in the on-demand hints, interaction was not beneficial.

Log data. We examined log data from the Cognitive Tutor for hints that were requested during “not given” steps (i.e., steps during which students must apply a geometry rule to calculate an answer). Separate two-way ANOVAs were conducted for percent-

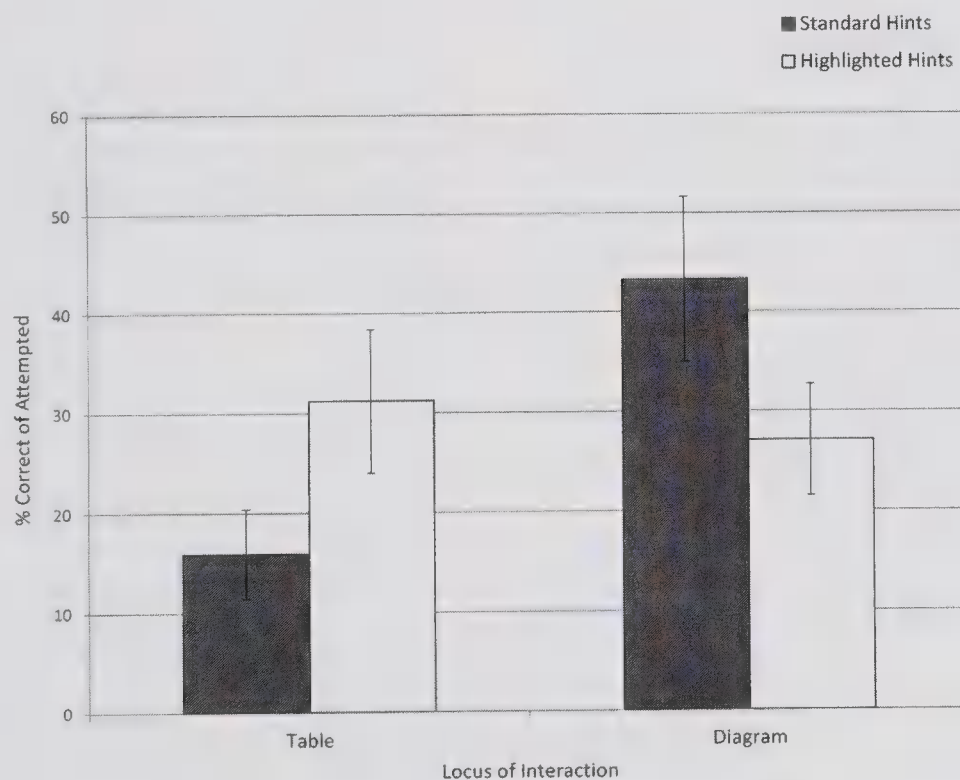


Figure 8. Performance on rule application items by experimental conditions. Error bars show the standard error of the mean.

age of steps on which a hint was requested and for the average amount of time spent on each hint. As can be seen in Table 3, there were no significant effect main effects or interactions ($F_s < 1$). According to log data, all conditions used hints in similar ways during intelligent tutoring practice.

Bivariate correlations were used to explore the potential relationship between hint use and eventual posttest performance. Since students who request many hints during intelligent tutoring likely are struggling with the content in general (Aleven & Koedinger, 2000), it is not surprising that increased use of tutor hints overall was negatively correlated with performance on assessment measures (see Table 4). However, an interesting pattern emerges when one considers individual conditions. For students who interacted with diagrams during intelligent tutoring, spending more time per highlighted hint was associated with better application of geometry rules (see Table 4); students who interacted with diagrams but saw standard hints did not show this pattern. For students who interacted with the solutions table and saw highlighted hints, spending more time per hint was associated with better performance on solvability judgments; students who interacted with tables but saw standard hints did not show this pattern. These data provide indirect evidence that visual representations of rule–diagram mappings may be useful, but only if students spend more time processing them. Conversely, these data also may indicate that some instructional scaffolds can reduce active processing: Some students may have used the highlighted hints as a shortcut to reduce the effort needed to process hint content.

Discussion

Results from Experiment 2 demonstrate that interacting with diagrams during intelligent tutoring may support spontaneous rule–diagram mappings that facilitate understanding of the geometry rules used during problem solving. Students who saw standard hints during intelligent tutoring were more successful in applying geometry rules to specific problems at posttest if they had interacted with the diagram than if they had used a solutions table. For students who saw highlighted hints, the locus of interaction did not affect their ability to apply geometry rules to problem-solving diagrams. Correlational results suggest that students who take the time to process visually represented rule–diagram mappings may develop better understanding of geometry rules, especially when they are interacting with diagram features during problem solving.

Why didn't providing students with visual representations of rule–diagram mappings support learning of geometry rules when students interacted with diagrams? One possibility is that students might have processed the highlights in shallow ways, for example, by attending to numerical quantities associated with the highlights rather than the visual features of the geometry diagrams themselves. Another possi-

bility is that providing rule–diagram mappings may reduce generative processes during problem-solving practice. When learners interact with diagrams during intelligent tutoring, they may be spontaneously making connections between text-based hints and visual problem features; providing highlighted hints that show these connections may negate such generative processing and reduce “desirable difficulty” (Bjork, 1994, 1999). This may have occurred even though a poststudy survey showed that all students reported similar approaches to the problem-solving task and similar (positive) reactions to the support that it provided (see the Appendix for a description of the survey and its analysis; survey items are provided in Table A1). Thus, students may need support in order to process rule–diagram mappings more deeply, potentially by engaging in constructive or interactive activities (Chi, 2009). Since increased time per hint provides circumstantial evidence of active learning processes (Shih, Koedinger, & Scheines, 2008), the correlational results support this possibility: Students who interacted with diagrams *and* engaged in active processing of the rule–diagram mappings (i.e., spent more time with the highlighted hints) given to them were better able to apply geometry rules to problem diagrams. Thus, a key question is whether scaffolding active processing of the rule–diagram mappings results in better understanding of geometry rules. Experiment 3 was designed to test this question by systematically varying whether visually based rule–diagram mappings were required and, if they were, by varying whether they were provided to students or generated by students.

Experiment 3

In this experiment, we examined the impact of providing students with rule–diagram mappings or requiring students to generate rule–diagram mappings using geometry problem-solving diagrams. In order to control for student attention to visual diagrams during problem solving, the locus of interaction was the geometry diagram for all conditions in this experiment (i.e., all students used the interactive diagram version of the Cognitive Tutor). The instantiation of rule–diagram mapping using highlighted diagram features was kept the same as in the previous experiment. In order to increase the frequency with which students encountered/generated these rule–diagram mappings, we modified the system so that the mappings occurred after each error in the Cognitive Tutor. Errors tend to be more frequent than hint use, although both tend to occur when students are working on steps for which they lack adequate knowledge.

Method

Participants. Participants were 83 students from five 10th grade geometry classrooms at the same vocational school as in Experiments 1 and 2 but in a different academic year (i.e., a unique

Table 3
Experiment 2: Log Data (Means and Standard Deviations) Associated With Hint Use During Intelligent Tutoring

Cognitive Tutor log measure	Table interaction		Diagram interaction	
	Standard hints (<i>n</i> = 17)	Highlighted hints (<i>n</i> = 19)	Standard hints (<i>n</i> = 20)	Highlighted hints (<i>n</i> = 11)
Average time (in seconds) per hint	14.0 (11.0)	16.1 (13.4)	12.5 (6.9)	13.7 (6.3)
% of steps for which a hint was requested	29.1 (11.1)	31.3 (9.4)	32.0 (10.7)	30.5 (9.8)

Table 4

Correlations Between Hint Use During Practice and Posttest Item Performance (% Correct of Attempted)

Condition	Solvability judgments	Numerical solutions	Rule application
Overall ($n = 44$)			
% steps for which hint was used	-.32*	-.41**	-.22
Average time (seconds) per hint	.07	.01	-.02
Highlighted hints, diagram interaction ($n = 8$)			
% steps for which hint was used	-.31	-.78*	-.74*
Average time (seconds) per hint	-.23	.38	.81*
Highlighted hints, table interaction ($n = 11$)			
% steps for which hint was used	-.13	<-.01	-.08
Average time (seconds) per hint	.61*	-.07	-.25
Standard hints, diagram interaction ($n = 10$)			
% steps for which hint was used	-.36	-.69*	.30
Average time (seconds) per hint	-.61	.20	.15
Standard hints, table interaction ($n = 15$)			
% steps for which hint was used	-.36	-.49	-.56*
Average time (seconds) per hint	.17	-.16	-.11

* $p \leq .05$. ** $p \leq .01$

group of students). All five classes were taught by the same teacher. Grade-matched triplets of students were identified within each class, using students' first semester geometry grades. From every grade-matched triplet, students were randomly assigned to one of three experimental conditions, described below.

Materials.

Geometry Cognitive Tutor. The Geometry Cognitive tutor interfaces varied only in whether or not rule-diagram mapping was enforced following students' errors and whether students generated or were provided with these mappings. The three conditions were as follows.

Student-generated mapping. The purpose of the student mapping condition was to investigate the impact of requiring students to generate visually based rule-diagram mappings. To generate these mappings, students selected the diagram features relevant to the specific geometry rules that were used to solve problem sub-goals in the Cognitive Tutor. Students selected diagram features by clicking on the relevant diagram elements in the tutor's interactive diagram; clicking an element highlighted it in the diagram (see Figure 9). If a student entered an incorrect answer or reason during practice, she or he was locked out of the numerical solution field until she or he identified a correct geometry rule to justify the problem-solving step. If the rule also was entered incorrectly, students were required to revise their entry until a correct rule had been identified. Once a correct geometry rule was entered, students were required to highlight the diagram elements relevant to that rule (see Figure 9), forming an integrated rule-diagram representation. The tutor scaffolded rule-diagram mappings by prompting students to highlight each diagrammatic feature that was necessary to apply a named geometry principle.

For example, in Figure 9, after making an error in calculating the measure for Angle ABC, the student has selected the "Interior Angles Same Side" principle. The tutor has generated answer fields for all the diagrammatic features necessary to apply that principle: parallel lines, a transversal (that cuts the parallel lines), and two angles that are created by the intersection of the transversal with the parallel lines. The student has selected the parallel

lines and the transversal in Figure 9a and must now select (i.e., click on) each relevant angle to complete the highlighting seen in Figure 9b. Students received immediate feedback on each highlighted feature. Incorrect highlights turned red in the diagram and the accompanying answer area. Students were required to revise incorrect highlights until a correct, highlighted representation was generated. Correct highlights remained visible until the problem-solving step was completed.

Tutor-provided mapping. This condition utilized the same visually based rule-diagram mappings as the student highlighting condition, but in this case the mappings were provided by the tutor. Following an error, students were required to identify the correct geometry rule before completing any other step. As in the student-generated mapping condition, students were required to revise their entries for the geometry rule until a correct rule had been identified. Once a student had identified the geometry rule that justified the problem-solving step, the tutor automatically highlighted the diagram features necessary to apply that rule. In order to remain consistent with the information in the student-generated mapping condition, the tutor provided a textual list (i.e., correctly completed answer fields) of the highlighted diagram features in the adjacent work area. Tutor-provided highlighting and the displayed answer fields were identical to the final student-generated highlighting in a problem-solving step (see Figure 9b).

No mapping (control). The control condition was the diagram interaction condition from Experiments 1 and 2. This condition did not involve any highlighting of visual diagram features by either students or the ITS. Students completed numerical solutions and selected geometry rules for each problem-solving step.

Assessments. Assessments followed the same format as in Experiment 1.

Statistical analyses. A series of three repeated-measures ANOVAs were conducted where the independent variable was mapping condition (student-generated mapping, tutor-provided mapping, or no mapping) and the repeated factor was test time (immediate posttest, delayed posttest). As in Experiments 1 and 2, a separate analysis was conducted for each of the assessment types

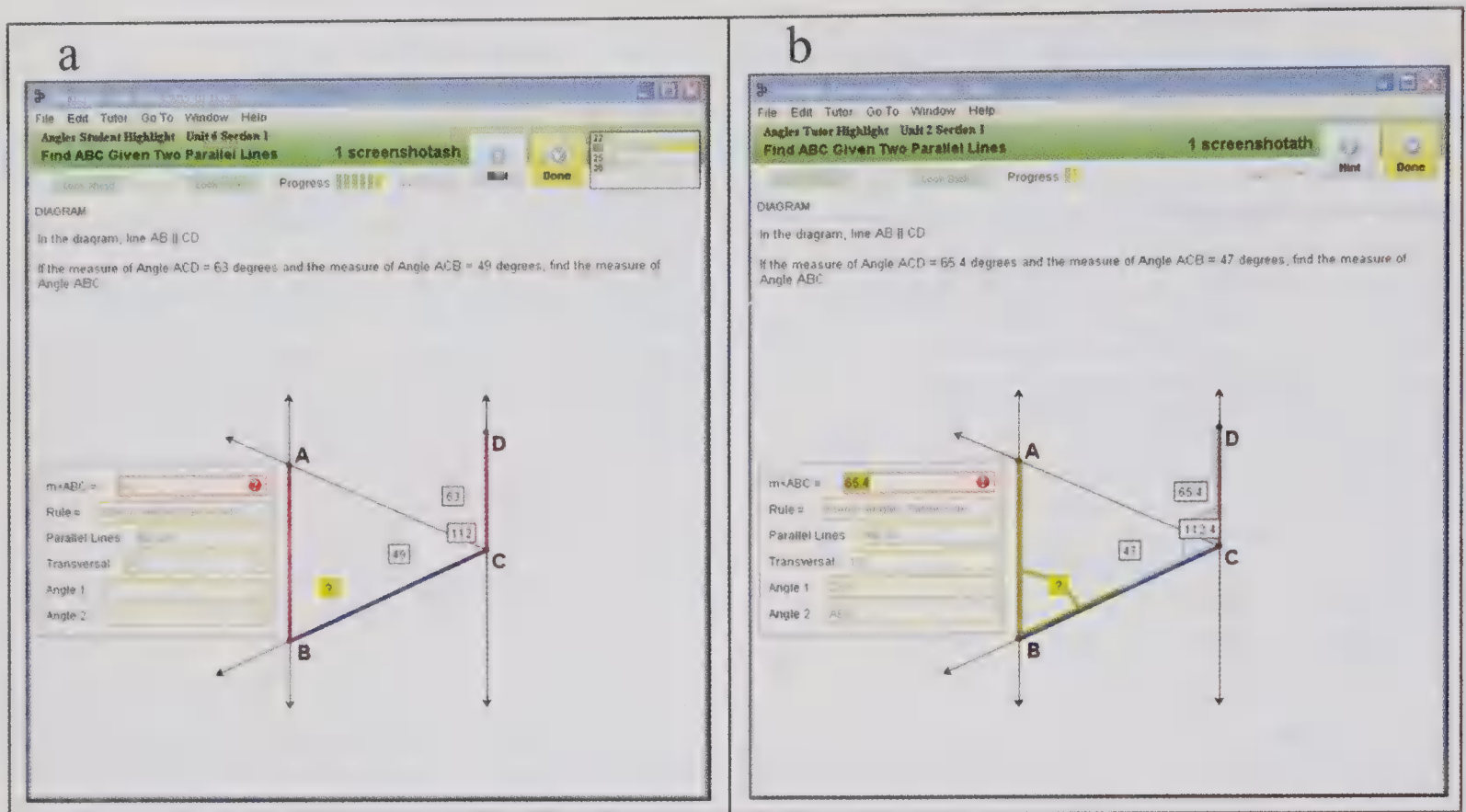


Figure 9. In-progress student highlighting of diagram (9a) and completed highlighting of diagram (9b) for the interior angles, same side principle.

(solvability decisions, numerical solutions, and rule application items), and alpha levels were set at .017 using a Bonferroni correction ($.05/3 = .017$).

Procedure. The procedure was the same as in Experiment 1 except that students were given up to 45 min to complete the posttest and 30 min to complete the delayed posttest.

Results

As a check of random assignment, a multivariate analysis of variance (MANOVA) was conducted for pretest scores where condition was the independent variable and percent correct of attempted items for solvability decisions, numerical solutions, and rule application were the dependent variables (all were within acceptable limits for kurtosis and skewness: between ± 2). There were no significant differences between conditions on pretest performance (numerical solutions: $F_{(2, 79)} = 1.0, p > .36$; all other item types: $F_s < 1$), as can be seen in Table 5. Thus, pretest data were not analyzed further.

Overall, 34 students were present for both the posttest and delayed posttest during the course of the study. Again, the attrition rate was high but comparable to that in other studies at the school (Salden et al., 2010; Walker et al., 2009) and comparable across conditions (see Table 5). Means and standard deviations are presented in Table 5.

Solvability decisions. There were no main effects of test time or condition ($F_s < 1$). The interaction between test time and condition did not reach statistical significance ($F_{(2, 31)} = 2.9, p = .07, \eta_p^2 = .16$).

Numerical solutions. There was no main effect of test time ($F_{(1, 31)} = 1.4, p = .24, \eta_p^2 = .04$) or condition ($F < 1$) and no interaction between test time and condition ($F_{(2, 31)} = 1.2, p = .31, \eta_p^2 = .07$).

Rule application. Although there were no main effects of test time or condition ($F_s < 1$), there was a significant test time by condition interaction ($F_{(2, 31)} = 4.7, p = .016, \eta_p^2 = .23$; see Figure 10). Tukey-Kramer post hoc comparisons showed a significant difference between student-generated mapping and tutor-provided mapping at delayed posttest ($p < .05$); the no-mapping condition fell between the other groups at delayed posttest and was not significantly different from either (see Table 5). There were no significant group differences at immediate posttest.

Discussion

Results show that student-generated rule–diagram mappings supported better long-term understanding of geometry rules as evidenced by students' abilities to apply domain principles (geometry rules) to specific problem representations (geometry diagrams) at delayed posttest. The same pattern was seen for students' performance on solvability decisions, although the effect did not reach the level of statistical significance. These results highlight a trade-off between generative processing and immediate outcomes. Requiring students to generate their own rule–diagram mappings during practice may have initially added demands that compromised immediate performance compared to the other conditions. However, significant differences

Table 5
Experiment 3 Means (and Standard Deviations) for Assessment Items

Item type	Student-generated mapping	Tutor-provided mapping	No mapping (control)
Pretest			
	(n = 28)	(n = 29)	(n = 25)
Solvability decisions			
% correct	57.8 (15.5)	55.4 (16.9)	58.3 (15.0)
% attempted	100 (0)	99.6 (2.3)	100 (0)
% correct of attempted	57.8 (15.5)	55.7 (3.2)	58.3 (15.0)
Numerical solutions			
% correct	28.6 (25.4)	21.3 (24.4)	25.7 (23.9)
% attempted	61.3 (38.6)	48.9 (30.4)	53.7 (32.5)
% correct of attempted	49.8 (33.8)	36.6 (34.4)	44.0 (35.5)
Rule application			
% correct	18.8 (12.7)	20.8 (16.8)	19.1 (12.7)
% attempted	99.9 (0.6)	99.5 (23.7)	99.9 (0.6)
% correct of attempted ^a	18.6 (2.4)	20.7 (16.6)	18.9 (12.3)
Posttest			
	(n = 10)	(n = 11)	(n = 13)
Solvability decisions			
% correct	42.5 (20.2)	51.7 (21.3)	59.6 (24.6)
% attempted	100 (0)	96.6 (11.3)	94.2 (20.8)
% correct of attempted	42.5 (20.2)	54.1 (21.8)	62.5 (20.4)
Numerical solutions			
% correct	21.7 (19.3)	28.8 (25.9)	39.1 (29.9)
% attempted	36.3 (22.2)	44.9 (19.7)	57.2 (33.2)
% correct of attempted	37.5 (31.6)	48.5 (28.7)	55.0 (34.6)
Rule application			
% correct	18.4 (13.7)	22.7 (18.5)	27.4 (17.2)
% attempted	99.7 (10.4)	96.6 (11.3)	94.0 (20.7)
% correct of attempted	18.5 (13.8)	23.4 (18.2)	27.9 (16.1)
Delayed posttest			
	(n = 10)	(n = 11)	(n = 13)
Solvability decisions			
% correct	57.5 (23.0)	52.3 (20.8)	46.2 (18.7)
% attempted	100 (0)	100 (0)	93.4 (21.7)
% correct of attempted	57.5 (23.0)	52.3 (20.8)	51.0 (24.7)
Numerical solutions			
% correct	26.3 (21.6)	20.5 (14.0)	21.2 (18.7)
% attempted	52.5 (20.2)	56.3 (23.0)	66.7 (27.4)
% correct of attempted	43.2 (28.8)	38.7 (22.1)	36.4 (28.3)
Rule application			
% correct	29.4 (12.2)	17.1 (14.3)	19.7 (16.3)
% attempted	100 (0)	100 (0)	93.8 (21.7)
% correct of attempted ^a	29.0 (12.0)	16.9 (14.2)	19.3 (16.0)

^a Percent correct of attempted can be lower than percent correct if students attempt to solve an unsolvable item.

emerged 1 month later, where students who generated rule-diagram mappings demonstrated the strongest performance on application items, especially compared to students who were provided with these mappings.

In this study, all conditions utilized the interactive diagram version of the Cognitive Tutor. Thus, all students were attending to and interacting with the diagram as they engaged in problem-solving practice and, as a consequence, could have engaged in some form of spontaneous rule-diagram mapping. In Experiment 3 we sought to determine if generating or providing visual representations of rule-diagram mappings across problems could improve student learning more than spontaneous mappings facilitated by interactive diagrams. Results show that students who were scaffolded in generating rule-diagram mappings gained benefits beyond those gained by students who

were provided with the rule-diagram mappings, even though those gains were not apparent at immediate posttest. The current findings demonstrate that visually based interactions can be used to support understanding of domain principles that justify problem-solving steps, especially at longer retention intervals. It should be noted that although the pattern of performance in Figure 10 may seem to suggest that students improved their understanding of geometry rules from posttest to delayed posttest, this likely is an artifact of the types of items on the immediate versus delayed posttest (where delayed posttest items included less complex diagrams). If less complex diagrams resulted in somewhat "easier" items at delayed posttest, it would be reasonable to observe increases in performance if students largely retained (rather than increased) their knowledge.

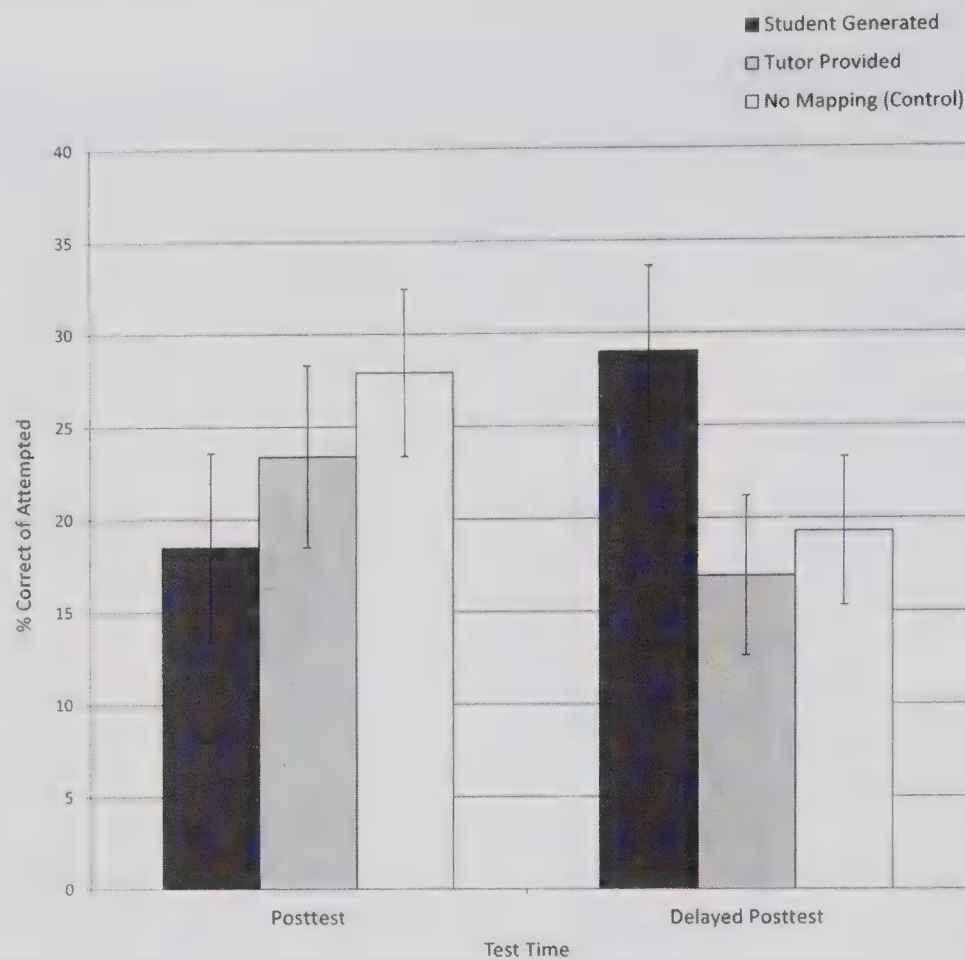


Figure 10. Performance on rule application items at posttest and delayed posttest by experimental condition. Error bars show the standard error of the mean.

General Discussion

Overall, this research demonstrates that advanced learning technologies can facilitate robust student learning via well-designed interactions that support students in making visually based mappings between domain principles and problem representations. Moreover, these studies demonstrate the potential impact of new learning technologies to use interactive visual components to support student learning. However, the research also demonstrates that interaction itself is not sufficient to increase learning: Interactions should be carefully designed to require students to generate connections between problem features and higher level domain principles. In this case, generating visual representations of the connection between geometry rules and diagram features, rather than having them provided, resulted in better long-term understanding of those rules.

Findings from Experiment 3 demonstrate the potential to design interactive elements that scaffold students' reasoning about specific problems in relation to domain knowledge. Learning about geometry rules was facilitated when students were scaffolded in generating representations (i.e., highlighted diagrams) that connected diagram features to geometry postulates and theorems. However, one might question the robustness of this finding since there were not significant differences between the student-generated mapping and the no-mapping conditions in Experiment 3. Why shouldn't scaffolding rule–diagram mapping be significantly better than no mapping? Most likely, some students in the no-mapping condition did not generate rule–diagram mapping

while others spontaneously engaged in such rule–diagram mapping. However, when the tutor *provided* the mappings, far fewer students engaged in generative processing. More simply, it is likely that no generative mapping was occurring in the tutor-provided mapping condition, some generative mapping may have been happening in the no-mapping (interactive diagram) condition, and generative mapping was *required* in the student-generated mapping condition. The degree to which relevant, generative activity is required by the tutoring interface predicts the pattern of performance on rule application items at delayed posttest (student-generated mapping > no mapping > tutor-provided mapping). However, it also should be noted that the tutoring interface strongly scaffolds rule–diagram mapping in this study (by prompting students to highlight each relevant diagram feature); thus, scaffolded reasoning steps may be contributing to the results independent of visually based generation. Future research is needed to explore the relative contributions of these factors.

The difficulty of designing additional interactions in a step-based ITS that result in further improvements to student learning has been documented by VanLehn (2011). VanLehn noted that the preponderance of evidence has demonstrated that attempts to scaffold student thinking to levels of granularity finer than step-based reasoning in ITS have not impacted student performance. That is, ITSs that require step-based and substep-based responses (where substep-based responses ask for reasoning behind a step) typically produce equivalent learning gains (VanLehn, 2011; VanLehn et al., 2007). In the current research, the lack of impact on problem-

solving success—specifically, performance on numerical solutions—is consistent with VanLehn’s (2011) findings. But the current research goes beyond previous findings to demonstrate that generative interactions that serve to connect problems and domain principles can facilitate longer lasting knowledge of the reasons that underlie problem-solving procedures, even when these interactions support substep-based reasoning. In the current research, steps required students to provide numerical answers and name relevant geometry rules, whereas substeps required students to reason about how selected rules function within a specific step. At the substep level, students mapped the connection between a domain-level rule (relevant to the step) and a specific problem diagram using more fine-grained reasoning (e.g., identifying the diagram features that were necessary to apply the rule in a particular step). Since even the control conditions in the current studies enforced accuracy at step levels (finding numerical answers and identifying geometry rules), we would not expect better rule–diagram mapping (at the substep level) to result in greater solution accuracy. However, we should expect rule–diagram mapping to help students achieve a deeper understanding of how and when specific geometry rules are used within problem-solving steps (i.e., improved performance on *rule application* items). Experiment 3 demonstrates that visual interactions can be an effective method to promote substep-level reasoning about domain rules, resulting in better long-term understanding of these rules. Findings from the current studies also demonstrate that substep interactions are sensitive to small changes in the focus of attention and in the amount of generative processing required by the interaction, making large effects difficult to achieve. Future research should continue to explore the potential for generative interactions to facilitate finer-grained reasoning, with specific exploration of visually based interactions in STEM areas that utilize diagrams or visual models.

The current results are consistent with prior research showing that students using an ITS can develop (shallow) problem-solving skills that allow them to be largely successful in reaching numerical solutions to problems without fully understanding the geometry rationale for these solutions (Aleven & Koedinger, 2002). A similar finding has been noted for an ITS in physics (Andes): Although research studies spanning 5 years found that students using the Andes system scored higher than control students on conceptual measures, accuracy of numerical answers never was affected by tutor use (VanLehn et al., 2005). The current findings confirm that making accurate connections to geometry rules is not always critical to determining a correct numerical answer and that effective interactions in an ITS should target the development of conceptual knowledge. In this work, scaffolding mapping between domain principles (i.e., geometry rules) and problem features (i.e., diagram configurations) helped students develop problem-solving skills that were more closely tied to domain knowledge and, accordingly, represent movement away from shallow problem-solving.

It should be noted that the current results do not show compelling evidence that deep learning was achieved; overall, students’ raw levels of performance were relatively low, and students chose to skip a fair number of problems during assessment. As noted in the Limitations section, this may be due to the specific (lower performing) student population that participated in this research; however, it also may suggest that additional interventions are

needed to move students more decisively toward a level of understanding that could be described as “deep” rather than simply as “less shallow.”

Results from the current studies extend previous findings showing that pre-integrated materials can be less useful than materials that require students to actively generate an integrated representation (Bodemer et al., 2005); however, our results go beyond previous findings to show that the benefits of interactivity are moderated by the focus of student attention during interactivity as well as the target of the interactive actions. Although both Experiments 1 and 3 required students to generate rule–diagram mappings, only Experiment 3 required students to form a *visual* representation of the rule–diagram mapping using elements in the problem diagrams. The explanations used in Experiment 1 (where students clicked on numerical quantities to “name” diagram elements) did not direct students’ attention to the key visual features involved in mapping and may have been redundant with problem-solving strategies. The interaction technique used in Experiment 3 avoids this shortcoming by comprehensively targeting the visual features relevant to a principle’s application but is still a lightweight form of interaction that avoids placing undue demands on students during problem-solving practice (i.e., students can easily select diagram elements by clicking).

Results from Experiment 2 mainly were consistent with previous research showing that visual cues can equalize the benefits of providing versus generating instructional explanations (de Koning, et al., 2010). In Experiment 2, students who did not receive visual mappings benefited from interacting directly with the diagrams (as opposed to a solutions table), likely because they engaged in some spontaneous processing of rule–diagram mappings. But students who were provided with visual representations of the rule–diagram mappings (in the form of highlighted hints) did not benefit from diagram interaction (i.e., locus of interaction did not affect rule application when students saw highlighted hints). Providing rule–diagram mappings for students who interact with problem diagrams may reduce the degree to which they actively generate rule–diagram mappings on their own, or it may facilitate or invite shallow strategies. Although we do not have direct evidence to distinguish between these possibilities, results from Experiment 3 demonstrate that the generation of representations that depict rule–diagram mappings is more effective than providing such representations. In Experiment 3, prompting students to generate visual representations of rule–diagram mappings, rather than providing students with these mappings, resulted in better understanding of geometry rules at delayed posttest. This result occurred even though content and timing of the representations were equivalent. Overall, the current findings extend previous work by demonstrating that providing visual cues can be effective when students are not already attending to relevant diagram features, but scaffolding student generation of domain-relevant representations is more effective for long-term learning.

The implications of the current findings may extend beyond geometry to other domains where varied visual representations are used to reason about domain concepts (e.g., chemistry). Overall, this research suggests that scaffolding visual interactions in problem diagrams can be an effective way to support students’ understanding of how domain concepts apply to specific problems.

especially when such interactions build a representation that demonstrates the domain-level concept(s) operating on the problem. In other domains, different interactions may be necessary to achieve meaningful coordination between visual features and domain principles. For example, visual representations of chemistry molecules may need to be rotated or labeled during problem solving (Stieff, Ryu, & Dixon, 2010). More research is needed to understand rule–diagram mapping and its influences on learning in other domains.

Finally, it is important to remember that the benefits of student-generated rule–diagram mappings were not evident until delayed posttest, which occurred a month following the immediate posttest. Thus, the current work argues for the importance of assessing at longer delays when attempting to evaluate the potential of instructional interventions.

Limitations

The current research is not without its limitations. First, students at the vocational school where the studies were conducted may represent a lower achieving population compared to students in traditional school settings. This may help account for the relatively low overall levels of performance (when one considers raw scores) and of the percentage of problems that students attempted to solve. Although the potential to support learning with lower achieving populations is not trivial, future research should explore the impact of rule-mapping interactions with typical student populations. Second, absenteeism was a recognized problem at this vocational school. Although the rates of absenteeism seen during this study were similar to those in other studies conducted at the school (Salden et al., 2010; Walker et al., 2009), a more sensitive picture of tutor impact may be seen in educational contexts with consistent attendance. Third, the tutor language in these studies could have been more closely aligned to the formal language of geometry. Using the label *rule* to refer to the geometry postulates and theorems that justified problem-solving steps reflects an informal use of language that may have complicated students' reasoning. This was especially true in Experiment 1, where "rules" were associated with "applied to" fields. Interpretation of the "applied to" label required students (a) to recognize that this field targeted known (or previously solved) information rather than the unknown quantity and (b) to (perhaps unnaturally) separate the known arguments (e.g., two angles) from the unknown quantity during application of a geometry postulate or theorem. Future research should more carefully align the language within the tutoring environment to the language of the domain (e.g., drawing upon language from geometry proofs to require a "reason" rather than a "rule"). Finally, the research was conducted during a focused set of instructional units in the geometry curriculum. It remains to be seen how longer term use of rule–diagram mappings in an ITS affects student outcomes. Scaffolding of students' rule–diagram mappings may need to be removed as students gain competence with the task, since other research has shown that fading instructional support (Atkinson et al., 2003; Salden, Aleven, Schwonke, & Renkl, 2008) can improve student learning and knowledge transfer.

Conclusions

The current research informs the development of advanced learning technologies for domains that include visual and verbal information sources. Our work demonstrates that designing student interactions that promote learning requires careful attention to the representational forms with which students may interact. Although connecting diagram elements to domain rules via student-generated highlights supported long-term learning about these rules (Experiment 3), making these same connections by interacting with solved quantities was ineffective (Experiment 1). Interacting directly with diagrams appeared to facilitate spontaneous processing of rule–diagram mappings, but providing visual representations of rule–diagram mappings negated the effects of interaction (Experiment 2). Providing visual representations of rule–diagram mappings was not as effective as scaffolding student generation of these mappings (Experiment 3). Together, these results demonstrate that attempts to explore the assistance dilemma (Koedinger & Aleven, 2007) may require attention not only to the *kinds* of information being provided or withheld but also to the *representational format* of the information and the *interactions* available to promote deeper processing of the representations. Overall, our findings show that there are potential benefits in learning technologies that facilitate student interaction with multimedia and visual representations, especially when these interactions focus student attention and processing on key domain concepts.

References

- Aleven, V., & Koedinger, K. R. (2000). Limitations of student control: Do students know when they need help? In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent tutoring systems: 5th International Conference, ITS 2000* (Lecture Notes in Computer Science, Vol. 1839, pp. 292–303). Berlin, Germany: Springer-Verlag. doi:10.1007/3-540-45108-0_33
- Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, 26(2), 147–179. doi:10.1207/s15516709cog2602_1
- Aleven, V., Koedinger, K. R., Sinclair, H. C., & Snyder, J. (1998). Combatting shallow learning in a tutor for geometry problem solving. In B. P. Goettl, H. M. Half, C. L. Redfield, & V. J. Shute (Eds.), *Intelligent tutoring systems: 4th International Conference, ITS '98*, (Lecture Notes in Computer Science, Vol. 1452, pp. 364–373). Berlin, Germany: Springer-Verlag. doi:10.1007/3-540-68716-5_42
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4(2), 167–207. doi:10.1207/s15327809jls0402_2
- Anderson, J. R., & Lebière, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.
- Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, 70(2), 181–214.
- Atkinson, R. K., & Renkl, A. (2007). Interactive example-based learning environments: Using interactive elements to encourage effective processing of worked examples. *Educational Psychology Review*, 19(3), 375–386. doi:10.1007/s10648-007-9055-2
- Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Combining fading with prompting fosters learning. *Journal of Educational Psychology*, 95, 774–783. doi:10.1037/0022-0663.95.4.774

- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.
- Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriati (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.
- Bodemer, D., Ploetzner, R., Bruchmüller, K., & Hacker, S. (2005). Supporting learning with interactive multimedia through active integration of representations. *Instructional Science*, 33, 73–95. doi:10.1007/s11251-004-7685-z
- Bodemer, D., Ploetzner, R., Feuerlein, I., & Spada, H. (2004). The active integration of information during learning with dynamic and interactive visualisations. *Learning and Instruction*, 14, 325–341. doi:10.1016/j.learninstruc.2004.06.006
- Butcher, K. R., & Aleven, V. (2007). Integrating visual and verbal knowledge during classroom learning with computer tutors. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 137–142). Austin, TX: Cognitive Science Society.
- Carnegie Learning. (2007). *Teachers' implementation guide: Volume 1*. Pittsburgh, PA: Author.
- Carnegie Learning. (2010). *Geometry* (2nd ed.). Pittsburgh, PA: Author.
- Carter, J. A., Cuevas, G. J., Day, R., Malloy, C., & Cummins, J. (2012). *Glencoe geometry teacher's edition*. New York, NY: McGraw Hill.
- Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, 1, 73–105. doi:10.1111/j.1756-8765.2008.01005.x
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145–182. doi:10.1207/s15516709cog1302_1
- Chi, M. T. H., de Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 439–477.
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152. doi:10.1207/s15516709cog0502_2
- Conati, C., & VanLehn, K. (2000). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education*, 11, 389–415.
- de Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2007). Attention cueing as a means to enhance learning from an animation. *Applied Cognitive Psychology*, 21(6), 731–746. doi:10.1002/acp.1346
- de Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2010). Learning by generating vs. receiving instructional explanations: Two approaches to enhance attention cueing in animations. *Computers & Education*, 55(2), 681–691. doi:10.1016/j.compedu.2010.02.027
- Glenberg, A. M., & McDaniel, M. A. (1992). Mental models, pictures, and text: Integration of spatial and verbal information. *Memory & Cognition*, 20(5), 458–460. doi:10.3758/BF03199578
- Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, 14(5), 462–466. doi:10.1111/1467-9280.02454
- Hausmann, R. G. M., & Chi, M. T. (2002). Can a computer interface support self-explaining? *Cognitive Technology*, 7(1), 4–14.
- Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32, 175–191. doi:10.1016/j.intell.2003.12.001
- Hilbert, T. S., Renkl, A., Kessler, S., & Reiss, K. (2008). Learning to prove in geometry: Learning from heuristic examples and how it can be supported. *Learning and Instruction*, 18, 54–65. doi:10.1016/j.learninstruc.2006.10.008
- Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3), 239–264. doi:10.1007/s10648-007-9049-0
- Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, 14(4), 511–550. doi:10.1207/s15516709cog1404_2
- Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30–43.
- Kozma, R. (2003). The material features of multiple representations and their cognitive and social affordances for science understanding. *Learning and Instruction*, 13, 205–226. doi:10.1016/S0959-4752(02)00021-X
- Lovett, M. C., & Anderson, J. R. (1994). Effects of solving related proofs on memory and transfer in geometry problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(2), 366–378. doi:10.1037/0278-7393.20.2.366
- Mayer, R. E. (2001). *Multimedia learning*. New York, NY: Cambridge University Press. doi:10.1017/CBO9781139164603
- Mayer, R. E., & Anderson, R. B. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology*, 84, 444–452. doi:10.1037/0022-0663.84.4.444
- Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, 91(2), 358–368. doi:10.1037/0022-0663.91.2.358
- Moreno, R., Ozogul, G., & Reisslein, M. (2011). Teaching with concrete and abstract visual representations: Effects on students' problem solving, problem representations, and learning perceptions. *Journal of Educational Psychology*, 103(1), 32–47. doi:10.1037/a0021995
- Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, 21, 1–29. doi:10.1207/s15516709cog2101_1
- Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: The effects of example variability and elicited self-explanations. *Contemporary Educational Psychology*, 23(1), 90–108. doi:10.1006/ceps.1997.0959
- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. T. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2), 249–255. doi:10.3758/BF03194060
- Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(3), 456–468. doi:10.1037/0278-7393.15.3.456
- Ryan, M. (2011). *Geometry essentials for dummies*. Hoboken, NJ: Wiley.
- Salden, R. J. C. M., Aleven, V., Schwonke, R., & Renkl, A. (2008). *Are worked examples and tutored problem solving synergistic forms of support?* In Proceedings of the 8th International Conference of the Learning Sciences (ICLS'08) (Vol. 3, pp. 119–120). Available from the International Society of the Learning Sciences website: <http://www.isls.org/icls2008/proceedings.html>
- Salden, R. J. C. M., Aleven, V., Schwonke, R., & Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. *Instructional Science*, 38(3), 289–307. doi:10.1007/s11251-009-9107-8
- Schworm, S., & Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology*, 99(2), 285–296. doi:10.1037/0022-0663.99.2.285
- Shih, B., Koedinger, K. R., & Scheines, R. (2008). A response time model for bottom-out hints as worked examples. In R. S. J. de Baker, T. Barnes, & J. Beck (Eds.), *Educational Data Mining 2008: 1st International Conference on Educational Data Mining, proceedings* (pp. 117–126). Retrieved from <http://repository.cmu.edu/philosophy/> 428

- Stieff, M., Hegarty, M., & Deslongchamps, G. (2011). Identifying representational competence with multi-representational displays. *Cognition and Instruction*, 29(1), 123–145. doi:10.1080/07370008.2010.507318
- Stieff, M., Ryu, M., & Dixon, B. (2010). Students' use of multiple strategies for spatial problem solving. In K. Gomez, L. Lyons, & J. Radinsky (Eds.), *Proceedings of the 9th International Conference of the Learning Sciences (ICLS '10)* (Vol. 1, pp. 765–772). Available from the International Society of the Learning Sciences website: http://www.isls.org/icls2010/conf_proceedings.html
- Tabbers, H. K., Martens, R. L., & van Merriënboer, J. J. (2004). Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British Journal of Educational Psychology*, 74, 71–81. doi:10.1348/000709904322848824
- Thomas, L. E., & Lleras, A. (2007). Moving eyes and moving thought: On the spatial compatibility between eye movements and cognition. *Psychonomic Bulletin & Review*, 14(4), 663–668. doi:10.3758/BF03196818
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. doi:10.1080/00461520.2011.611369
- VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, 31, 3–62. doi:10.1080/03640210709336984
- VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., . . . Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, 15, 147–204.
- Walker, E., Rummel, N., & Koedinger, K. R. (2009). Integrating collaboration and intelligent tutoring data in the evaluation of a reciprocal peer tutoring environment. *Research and Practice in Technology Enhanced Learning*, 4(3), 221–251. doi:10.1142/S179320680900074X
- Wilkin, B. (1997). *Learning from explanations: Diagrams can "inhibit" the self-explanation effect* (AAAI Technical Report FS-97-03). Retrieved from <http://aaai.org/Papers/Symposia/Fall/1997/FS-97-03/FS97-03-017.pdf>

Appendix

Experiment 2 Poststudy Survey Description and Analysis

A 14-item Likert-style survey was used to gauge students' reactions to the intelligent tutoring system using a 6-point scale (from 1 = *totally disagree* to 6 = *totally agree*). Multivariate results showed no significant main effect of the locus of interaction ($F_{(13, 40)} = 1.1, p > .40$) or tutor format ($F < 1$) and no interaction among the independent variables ($F_{(13, 40)} = 1.3, p > .27$). The first item ("I am color-blind . . .") was dropped from this analysis based upon kurtosis and skewness; all other questions were within acceptable limits for kurtosis and skewness (between ± 2).

(Appendix continues)

Table A1

Experiment 2 Poststudy Survey Items: Means (and Standard Deviations) for Responses by Condition

Item (1 = <i>totally disagree</i> , 6 = <i>totally agree</i>)	Table interaction		Diagram interaction	
	Standard hints (n = 15)	Highlighted hints (n = 15)	Standard hints (n = 17)	Highlighted hints (n = 9)
1. I am color-blind or I have difficulty seeing colors.	1.0 (0)	1.5 (1.4)	1.2 (0.6)	1.0 (0)
2. I thought that the tutor was easy to use.	4.0 (1.3)	3.9 (1.4)	4.4 (1.1)	4.1 (0.6)
3. It was easy to lose track of what angle I was working on while using the tutor.	2.7 (1.2)	4.1 (1.5)	3.4 (1.2)	3.1 (1.3)
4. The question marks in the images helped me figure out what I needed to solve.	3.1 (1.8)	4.2 (1.5)	3.8 (2.0)	4.3 (1.3)
5. It was easy to figure out the order in which I needed to find the answers in each problem.	4.3 (1.3)	3.5 (1.5)	2.9 (1.7)	3.3 (1.4)
6. I paid a lot of attention to the diagram while I was solving problems.	4.0 (1.5)	4.2 (1.1)	4.1 (1.3)	4.2 (0.8)
7. It was easy to understand the hints that I got on the computer.	4.3 (2.0)	3.1 (1.3)	3.7 (1.3)	4.0 (1.2)
8. The hints made the diagrams easier to understand.	4.5 (1.5)	3.7 (1.2)	3.8 (1.1)	4.3 (1.4)
9. I used the hints more during this study than I normally do when using the Carnegie Learning tutor.	4.4 (1.5)	4.4 (1.8)	4.1 (1.5)	3.8 (1.6)
10. The geometry rules in the glossary were easy to understand.	4.5 (1.4)	3.9 (1.3)	4.0 (1.2)	3.9 (1.3)
11. It is better to ask my teacher for help than to ask for a hint on the computer.	2.9 (1.8)	3.5 (1.5)	2.9 (1.9)	1.8 (1.2)
12. I think I learned a lot by using the tutor.	3.7 (1.7)	3.7 (1.4)	3.8 (1.1)	3.8 (1.0)
13. The problems on paper (the posttest) were harder than the problems in the tutor.	4.3 (1.9)	4.2 (1.6)	3.7 (1.9)	4.0 (2.0)
14. I liked participating in this study.	3.0 (1.6)	2.8 (1.9)	3.7 (1.8)	3.6 (1.2)

Received December 16, 2011

Revision received December 11, 2012

Accepted December 13, 2012 ■

Writing Pal: Feasibility of an Intelligent Writing Strategy Tutor in the High School Classroom

Rod D. Roscoe and Danielle S. McNamara
Arizona State University

The Writing Pal (W-Pal) is a novel intelligent tutoring system (ITS) that offers writing strategy instruction, game-based practice, essay writing practice, and formative feedback to developing writers. Compared to more tractable and constrained learning domains for ITS, writing is an ill-defined domain because the features of effective writing are difficult to quantify and individual writers can employ diverse strategies to achieve similar goals. The development of an ITS in an ill-defined domain presents particular challenges regarding comprehensive instruction, modularized content, extended practice, and formative feedback. In this article, we describe how the development of W-Pal has uniquely addressed these concerns and present the results of a study assessing the feasibility of this system in high school English classrooms. This study included 2 teachers and their 141 10th grade English class students who utilized W-Pal over a 6-month period during the academic year. Log-file analyses showed that students used all aspects of W-Pal, but activity and engagement was uneven throughout the year and decreased over time. Essay scores improved over time and surveys indicated that students perceived the lessons, games, and feedback as beneficial. However, specific aspects of the learning environment were critiqued as annoying, challenging, or lacking specificity. Overall, the results suggest that the system was generally well-received by the students but also offer insights for the development of ITSs in ill-defined domains.

Keywords: intelligent tutoring systems, writing instruction, usability and feasibility testing, ill-defined learning domains

Intelligent tutoring systems (ITSs) provide adaptive, interactive, computer-based support for learning based on sound pedagogical principles (Graesser, McNamara, & VanLehn, 2005), and educators now have access to effective intelligent tutors in domains such as mathematics (Beal, Arroyo, Cohen, & Woolf, 2010), geometry (Aleven & Koedinger, 2002), biology (Michael, Rovick, Glass, Zhou, & Evens, 2003), physics (Graesser et al., 2004; VanLehn et al., 2005), computer literacy (Graesser et al., 2004), reading comprehension (McNamara, O'Reilly, Best, & Ozuru, 2006), and foreign language (Gamper & Knapp, 2002; Johnson & Wu, 2008). In this study, we examine the Writing Pal (W-Pal), an ITS that offers *writing strategy instruction* along with game-based practice, essay writing practice, and formative feedback to high school students. Historically, ITS development has focused on well-defined learning domains, in which fundamental concepts, procedures, and evaluation criteria are relatively constrained. In contrast, writing is an *ill-defined learning domain* because the features

of skilled writing are difficult to quantify, and individual writers may employ diverse strategies to achieve similar goals.

A particular focus of this study is how high school students perceive intelligent tutoring of writing in the classroom (Grimes & Warschauer, 2010). For ill-defined domains, in which evaluations of students' work are inherently debatable, such subjective reactions are crucial. Students who rebuff the ITS are unlikely to engage with the system over meaningful periods of instruction (i.e., several weeks, a semester, or a school year). Thus, we assume that feasibility depends upon whether the system is perceived as valid and valuable. At this stage in W-Pal's development, an experimental test of instructional efficacy was not warranted. Rather, it was most important for us to examine a) how and whether students use the W-Pal over time and b) students' perceptions of the utility and design of W-Pal. These data help to define the feasibility of the system and inform later development and deployment.

Computer Support for Writing Instruction

Several technologies have been developed to support students' writing by grading essays (Grimes & Warschauer, 2010; Shermis & Burstein, 2003), teaching summarization (Kintsch, Caccamise, Franzke, Johnson, & Dooley, 2007) and argumentation skills (Wolfe, Britt, Petrovich, Albrecht, & Kopp, 2009), or scaffolding essay composition (Proske, Narciss, & McNamara, 2012; Rowley & Meyer, 2003). An important question is how well technologies address the pedagogical needs arising from the ill-defined nature of writing. Ill-structured problems possess ambiguous goals, solution paths, or assessment criteria (Simon, 1973). Lynch, Ashley, Pinkwart, and Aleven (2009, p. 258) argued that learning domains

This article was published Online First September 9, 2013.

Rod D. Roscoe and Danielle S. McNamara, Learning Sciences Institute, Arizona State University.

The research reported here was supported by Institute of Education Sciences, U.S. Department of Education Grant R305A080589 to Arizona State University. The opinions expressed are those of the authors and do not necessarily represent views of the institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Rod D. Roscoe, Learning Sciences Institute, Arizona State University, PO Box 872111, Tempe, AZ 85287-2111. E-mail: rod.roscoe@asu.edu

are ill-defined when “essential concepts, relationships, and procedures for the domain” and the “means to validate problem solutions or cases” are not specified by a single strong domain theory. There may be multiple conceptualizations of key problems and tasks and there may be multiple approaches for solving those problems. Given such ambiguity, assessment of solutions may also be context-dependent and subjective. Thus, ITSs in ill-defined domains must not only address the challenges common to any educational technology, but must also overcome unique hurdles that arise when appropriate content, tasks, evaluation, and feedback are uncertain.

The ill-defined nature of writing emerges from the many non-linear and interactive tasks that comprise the writing process (Deane et al., 2008; Flower & Hayes, 1981). For example, *pre-writing* involves generating and organizing ideas prior to writing, and *drafting* involves translating initial ideas and plans into coherent text. In persuasive writing, writers must frame their arguments precisely and objectively and support arguments with factual evidence. Subsequently, *revising* entails elaborating and reorganizing the text to improve overall quality. Throughout these stages, writers also develop cohesion, style, voice, and other global qualities. To help students navigate these complex demands, writing pedagogy emphasizes the importance of *strategy instruction* that equips students with (a) concrete strategies for diverse writing processes, (b) background knowledge for using the strategies, and (c) opportunities for extended practice (Graham, McKeown, Kiu-hara, & Harris, 2012; Graham & Perin, 2007). Effective interventions teach explicit strategies for planning, drafting, editing, and summarizing, along with information about how and why the strategies should be used (De la Paz & Graham, 2002).

Another aspect of the ill-defined nature of writing is the subjectivity of evaluation. Every essay exhibits unique content and errors that represent individual students' writing processes. To assign a score, essay graders (e.g., teachers) must interpret the appropriateness of these decisions in the context of the assignment. Writing assessment research has found this process to be challenging (Huot, 1996; Meadows & Billington, 2005). Over time and multiple instances of grading, human graders are unlikely to assign the same grades to the same essays consistently unless carefully trained to do so (Crossley & McNamara, 2011; Meadows & Billington, 2005). Such subjectivity also raises questions about how to give meaningful feedback. Research has emphasized the importance individualized, formative feedback that describes clear methods for improvement (McGarrell & Verbeem, 2007; Shute, 2008), such as strategies for developing arguments and evidence. In contrast to summative feedback on overall performance, formative feedback supports writing proficiency by making the means of progress explicit.

An analysis of writing instruction from the perspective of ill-defined learning domains thus suggests several design principles that are germane to any writing ITS. An intelligent writing tutor may need to combine (a) comprehensive strategy instruction across multiple phases of writing, (b) modularized content to accommodate different pedagogies or student needs, (c) opportunities for extended and varied writing practice, and (d) formative writing feedback related to writing proficiency and strategies. In the following sections, we consider how prior technologies have addressed these issues, and then discuss how these design principles have been uniquely implemented within the W-Pal tutoring system.

Automated Essay Scoring and Writing Evaluation

A significant challenge for computer-based writing instruction is the automated assessment of student writing and delivery of meaningful feedback. One advantage is that computer-based tools can evaluate many text features consistently and simultaneously, and apply the same criteria to all essays reliably and objectively. Indeed, automated essay scoring (AES) systems have been developed to facilitate essay grading using statistical modeling, machine learning, natural language processing (NLP), and latent semantic analysis (LSA). Prominent systems include *e-rater* (Attali & Burstein, 2006), *IntelliMetric* (Rudner, Garcia, & Welch, 2006), and *Intelligent Essay Assessor* (IEA; Landauer, Laham, & Foltz, 2003). Overall, AES scoring tends to be accurate. Human and computer-assigned scores correlate around .80 to .85 (Warschauer & Ware, 2006), with 40–60% perfect agreement (exact match of human and computer scores) and 90–100% adjacent agreement (human and computer scores within 1 point; e.g., Attali & Burstein, 2006; Dikli, 2006; Rudner et al., 2006). Over time, AES systems have become embedded within automated writing evaluation (AWE) systems that assign scores along with feedback on errors (e.g., spelling) and may include instructional scaffolds and learning management tools (Grimes & Warschauer, 2010). Examples include *Criterion* (e-rater scoring engine) from the Educational Testing Service (Burstein, Chodorow, & Leacock, 2004), *MyAccess* (IntelliMetric engine) from Vantage Learning (Grimes & Warschauer, 2010), and *WriteToLearn* (IEA engine) from Pearson Education (Landauer, Lochbaum, & Dooley, 2009).

Evaluations of AWE technologies have focused primarily on scoring accuracy, although a few studies have examined instructional efficacy. For example, Shermis, Burstein, and Bliss (2004) examined essay scores for over 1000 high school students, half of whom participated in typical classroom instruction and half of whom used *Criterion*. The two groups did not differ in holistic essay quality, although *Criterion* users produced longer essays with fewer mechanical errors. Rock (2007) obtained comparable results in a study with over 1,400 ninth grade students using *Criterion*. Finally, Kellogg, Whiteford, and Quinlan (2010) experimentally manipulated how much feedback 59 undergraduates received from *Criterion* on three essays. Students received feedback on all essays, one essay, or none. Holistic essay quality did not differ across conditions, although students who received more feedback displayed fewer mechanical errors in their essay revisions. In sum, *Criterion*¹ has been successful in improving student essays but primarily for mechanical properties, rather than holistic quality.

Grimes and Warschauer (e.g., Grimes & Warschauer, 2010; Warschauer & Grimes, 2008) have argued for the need to examine users' perceptions of AWE tools in the classroom. Successful deployment of writing technologies may depend upon whether teachers and students view the tools as valid, useful, and usable. Within this framework, Warschauer and Grimes (2008) examined perceptions of *Criterion* or *MyAccess* in four schools, obtaining survey and interview data from principals, teachers, and students (sixth to 12th grade). Both systems were perceived to increase students' motivation to write and improve writing quality, but the tools were used infrequently due to curricular conflicts. Students

¹ A literature search did not reveal similar evaluations of other systems.

did not always have time for extra writing assignments and the systems could not support every writing genre that teachers wished to cover. In addition, although the systems seemed to promote essay revising, most revisions focused on mechanics rather than content, organization, or style.

Grimes and Warschauer (2010) later examined MyAccess over a 3-year period in four middle schools. System use was initially infrequent—teachers did not create assignments in the system and students rarely revised. However, use increased over time as teachers became more comfortable with the technology. Survey data revealed both positive attitudes and skepticism. Teachers felt that MyAccess saved time, made teaching easier and more enjoyable, and allowed them to focus on higher level concepts. Teachers also reported that students were more motivated to write. However, teachers doubted the accuracy of the automated scores. They also favored MyAccess for persuasive essay writing but preferred traditional methods for informative, narrative, or analytical genre writing. Teachers also felt that MyAccess was suited to teaching sentence fluency and conventions, but less helpful for covering ideas, organization, voice, and word choice. Similarly, students perceived the system as usable and enjoyable, and felt that it increased their confidence and quantity of writing. However, students had trouble understanding the feedback and felt overwhelmed by the quantity of feedback. Some teachers had to create handouts to help students navigate the “pages of suggestions” from the system. In addition, some students began to focus on improving their scores rather than communicating their ideas.

In sum, research on AWE tools is promising but highlights how efficacy may be hindered by student and teacher perceptions. When users doubt the automated scores or feedback, or find them overwhelming, it is unlikely that the system will achieve its true potential. Another concern may be an emphasis on practice and feedback with less attention paid to strategy instruction or modular design. The fundamental purpose of AWE systems is the facilitation of writing assessment rather than teaching students about writing principles, goals, and strategies. Without such instruction, students may not be prepared to utilize the detailed writing feedback these tools offer. Last, an emphasis on error feedback may not satisfy the principle of formative feedback.

Computer-Based Tutorials for Writing

A few technologies have been created to teach specific writing skills or to scaffold the writing process. For example, the LSA-based *Summary Street* (Caccamise, Franzke, Eckhoff, Kintsch, & Kintsch, 2007; Kintsch et al., 2007) supports students' summarization skills. When students write summaries in the system, they receive graphical feedback showing how well their text captures the source materials. Research with *Summary Street* has shown that students wrote more effective summaries and spent more time engaged in writing when using the system. Perceptions of the system were also positive: students found the system easy to use and appreciated receiving feedback related to what they needed to fix in their summaries. Similarly, Wolfe et al. (2009) developed a web-based tutor for developing argument, counterarguments, and rebuttals. Evaluations of this system have shown the tutorial instruction improved students' ability to perform these tasks. Overall, such research suggests that computer-based tutorials can be

effective for training students on specific strategies related to writing.

Another technology, *Computer Tutor for Writing* (CTW; Rowley & Meyer, 2003) adopted a scaffolding approach in which students wrote essays in an enhanced word processor. The interface provided “workspaces” in which students could view descriptions, examples, and hints related to the writing process, such as goal-setting, drafting, and publishing. A tracking system monitored completion of these tasks. Importantly, CTW did not provide a holistic score for essays, nor were students given error feedback or strategy guidance for improving their essays. Thus, writing support in CTW was instantiated solely as structured guidance during composition. An evaluation of the CTW with 471 middle and high school students (Rowley & Meyer, 2003) revealed no difference between control (i.e., no CTW training, $n = 174$) and experimental conditions (i.e., training with CTW, $n = 298$). Neither group improved from pretest to posttest with regards to essay scores; control participants' scores decreased by about 1%, whereas experimental participants' scores increased by about 2%.

Proske et al. (2012) adopted a similar scaffolding approach with the *escribo* system. In *escribo*, students receive online support for prewriting, drafting, and revising processes, along with feedback about their choices at each stage. Forty-two German university students practiced writing with or without the system in one training session and then wrote an unsupported essay in a posttest session. Overall, students who interacted with *escribo* spent more time planning their essays, which facilitated faster drafting of the text. *escribo* students also spent more time revising their essays and the resulting texts were rated as more comprehensible. Thus, when students are provided with both comprehensive strategy help and informative feedback on their writing process, computer-based tutorials for writing are more effective.

In sum, previous computer-based writing tutors have shown mixed results, which may be attributed to whether feedback was provided. Successful tutors for summarization and argumentation focused on fewer skills but offered feedback on students' performance. The main drawback is potentially their scope; they do not provide comprehensive or modular instruction related to the entire writing process. In contrast, CTW addressed all phases of writing with support for each task, but students did not receive strategy feedback. The system appeared to be of little benefit. However, when structured writing support is combined with feedback, as in *escribo*, empirical evidence suggests that a scaffolding approach can be effective.

The Writing Pal

In the development of W-Pal, we have sought to synthesize key principles of strategy instruction, modularity, extended practice, and formative feedback (McNamara et al., 2011). The interdisciplinary development of the initial version of W-Pal spanned over 3 years with input from cognitive psychology, linguistics, computer science, and English education.

Writing Strategy Modules

The principles of *comprehensive strategy instruction* and *modularized content* were instantiated in W-Pal via nine *Writing Strategy Modules* (see Table 1). The content for these modules were

Table 1
Summary of Strategy Training Module Content and Practice Games

Module	Description of Strategies	Practice Games
Prologue	Introduces W-Pal, the animated characters, and discusses the importance of writing	
Prewriting Phase		
Freewriting	Covers <i>freewriting</i> strategies for quickly generating essay ideas, arguments, and evidence prior to writing (<i>FAST PACE mnemonic</i>)	Freewrite Feud Freewrite Fill-In
Planning	Covers <i>outlining</i> and <i>graphic organizer</i> strategies for organizing arguments and evidence in an essay	Mastermind Outline Planning Pump
Drafting Phase		
Introduction Building	Covers strategies for writing introduction paragraph <i>thesis statements</i> , <i>argument previews</i> , and <i>attention-grabbing techniques</i> (<i>TAG mnemonic</i>)	Essay Launcher Dungeon Escape
Body Building	Covers strategies for writing <i>topic sentences</i> and providing objective <i>supporting evidence</i> (<i>KISS & Tell mnemonic</i>)	Fix It – Introductions RAM-5
Conclusion Building	Covers strategies for restating the thesis, summarizing arguments, closing an essay, and maintain reader interest in conclusion paragraphs (<i>RECAP mnemonic</i>)	Fix It – Bodies Fix It – Conclusions Dungeon Escape
Revising Phase		
Paraphrasing	Covers strategies for expressing ideas with more <i>precise and varied wording</i> , <i>varied sentence structure</i> , and <i>condensing</i> choppy sentences	Adventurer's Loot Map Conquest
Cohesion Building	Covers strategies for adding cohesive cues to text, such as <i>connective phrases</i> , <i>clarifying undefined referents</i> , and <i>threading ideas</i> throughout the text	CON-Artist Undefined & Mined
Revising	Covers strategies for reviewing an essay for completeness and clarity (<i>TETRIS mnemonic</i>), and strategies for how to improve an essay by adding, removing, moving, or substituting ideas (<i>ARMS mnemonic</i>)	Speech Writer

developed based on research on writing strategy instruction (e.g., Graham & Perin, 2007) and substantive, iterative input from expert writing educators (Roscoe, Varner, Weston, Crossley, & McNamara, in press). Writing strategies were discussed by three animated agents via lesson videos (15–30 min each). Dr. Julie (teacher agent) explained the strategies, and Mike and Sheila (student agents) demonstrated them (Figure 1). These characters were developed using Media Semantics Character Builder software and text-to-speech voices by Loquendo. For many lessons, multiple strategies were organized by acronymic mnemonic devices, which can facilitate adolescent students' recall and use of writing strategies (e.g., De la Paz & Graham, 2002). Quiz and game-like checkpoints were embedded in the lessons to reinforce

the content, and students could take notes. All modules were accessible from a "Lessons Tab" in the W-Pal interface, which allowed users to progress through the modules in a flexible order.

Game-Based Practice

The principle of *opportunities for extended feedback* was realized by developing two broad modes of practice: *game-based practice* and *essay writing practice*. In W-Pal, a suite of educational games allows students to practice specific strategies outside of the context of complete essays. For example, students can practice strategies for evaluating evidence or building cohesion before applying these strategies in their own persuasive essays. Game-based practice was also chosen to address problems of student engagement. One challenge for ITSs is that students become bored and frustrated with extended practice (Bell & McNamara, 2007; Jackson & McNamara, 2013). Games offer a means of improving students' motivation to participate by leveraging their intrinsic enjoyment of gaming (Shank & Neeman, 2001).

In W-Pal, each Writing Strategy Module was associated with one or more practice games that students "unlock" by completing the lessons (see Table 1). This version offered 15 unique games. These games were iteratively developed by selecting key strategies covered in the lessons and then constructing *generative* or *identification* practice tasks. In generative practice, students write short texts (e.g., a conclusion paragraph) while applying one or more strategies. In identification practice, students examine text excerpts to label the strategies used, or to identify how strategies may be used to improve the text. These practice tasks were then embedded in diverse game mechanics and narratives. Feedback in the practice games was contextualized via the game design, such as winning or losing, earning points, the amount of fuel consumed by a spaceship, or the quality of treasure obtained. Thus, students could judge whether their strategy application was effective based on their

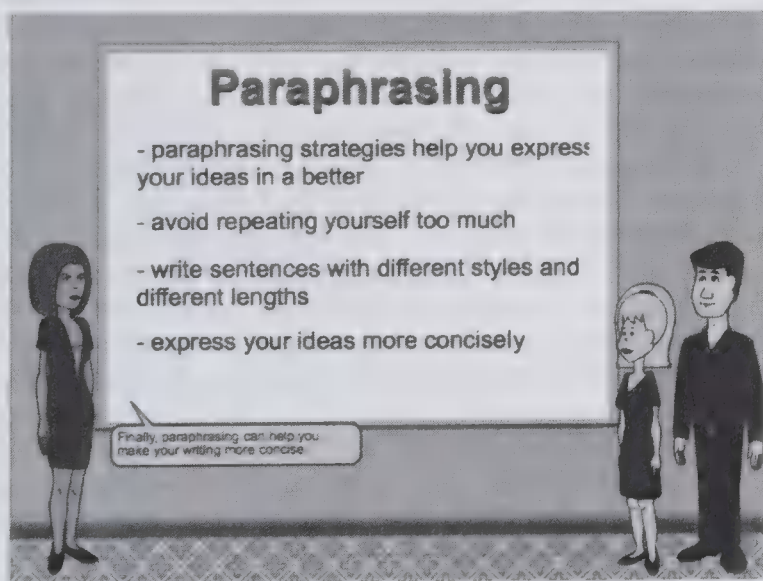


Figure 1. Screenshot of Writing Pal virtual classroom (Paraphrasing lesson).

game progress. In some cases, formative feedback was also offered, such as tips for succeeding in the game by using certain strategies or mnemonics.

To provide examples, we briefly present two games: Freewrite Feud and Essay Launcher. In *Freewrite Feud* (see Figure 2), students were given several minutes to freewrite on a persuasive writing prompt. For each prompt, a hidden list of key words and concepts was constructed based on a previous corpus of freewrites. Students earned points by typing their ideas quickly and continuously, and earned additional points when their freewrites incorporated up to six of the key words. Because these key words were hidden from the player, this generative game encouraged students to practice brainstorming many ideas, arguments, and potential pieces of evidence because doing so would trigger the key words and earn a higher score.

Essay Launcher was an Introduction Building game (see Figure 3). In this identification game, students attempted to repair and rescue several spaceships. To “repair the ship,” students chose a thesis statement for an example introduction paragraph from a list of three options. To “set the course,” students turned a dial labeled with attention-grabbing techniques to identify the technique used in the paragraph. Once both selections were made, students consumed one fuel unit to launch the ship. If either choice was incorrect, the launch malfunctioned. Students then received feedback about introduction strategies and could try again. Points were based on rescued ships and remaining fuel. This game allowed students to practice evaluating key characteristics of essay introductions.

Essay-Based Practice and Feedback

The principles of *formative feedback* and *opportunities for extended practice* were supported by the W-Pal *Essay Writing Interface* (see Figure 4). W-Pal allowed students to practice writing timed persuasive essays using SAT-style prompts in which they could synthesize and apply strategies covered in any module. Students could select the prompt, set the time limit, and use a scratchpad for prewriting. Essays were written using a simple word processor and then submitted for automated assessment.

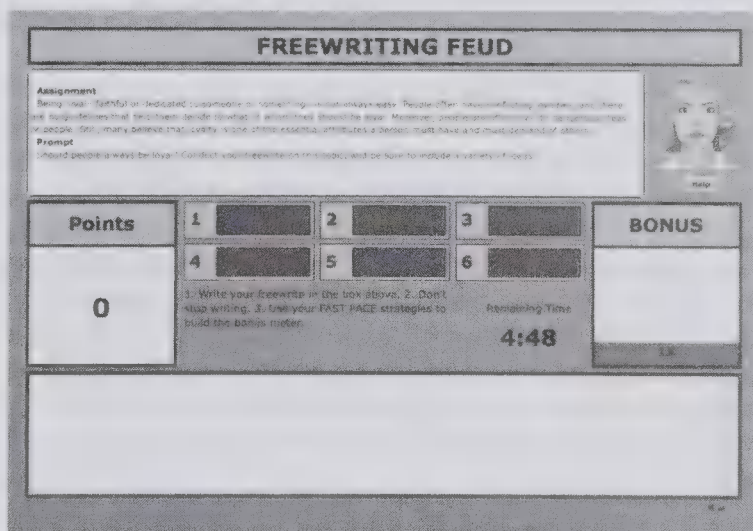


Figure 2. Freewriting Feud practice game (Freewriting).

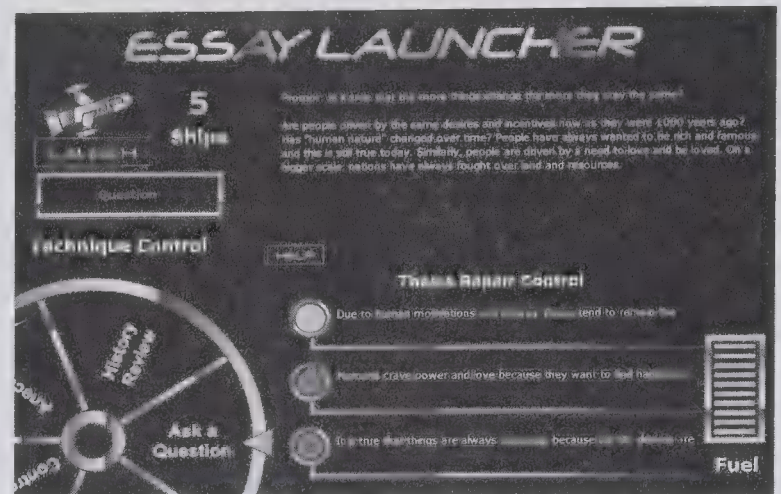


Figure 3. Essay Launcher practice game (Introduction Building).

W-Pal scoring is powered by NLP algorithms utilizing Coh-Metrix and other text analysis tools (Crossley & McNamara, 2011; Graesser & McNamara, 2012; McNamara, Crossley, & McCarthy, 2010; McNamara, Crossley, & Roscoe, 2012), and such algorithms are a key source of the *intelligence* of a writing ITS. Within ITSs that accept natural language as input (e.g., essays or verbal explanations of scientific processes), students' responses are open-ended and potentially ambiguous. When a user enters natural language into a system and expects useful and intelligent responses, NLP is necessary to interpret that input (McNamara, Crossley, & Roscoe, in press). In service to these goals, W-Pal utilizes Coh-Metrix to analyze text on several dimensions of cohesion including co-referential cohesion, causal cohesion, density of connectives, lexical diversity, temporal cohesion, spatial cohesion, and LSA. Coh-Metrix also calculates syntactic complexity and provides psycholinguistic data about words (parts-of-speech, frequency, concreteness, imaginability, meaningfulness, familiarity, polysemy, and hypernymy).

Essays submitted to W-Pal initially received a holistic rating from *poor* to *great* (6-point scale). Writers also received feedback that addressed particular writing goals and strategy-based solutions (see Figure 5). Such feedback was implemented as a series of scaffolded, threshold-based algorithms based on different linguistic properties and categories: *legitimacy* (e.g., proportion of non-words), *length* (e.g., number of words), *relevance* (e.g., occurrence of key words), and *structure* (e.g., number of paragraphs). For example, writers whose essays lacked elaboration (i.e., short essays) might receive feedback such as, “One way to expand your essay is to add additional relevant examples and evidence,” and prompts such as, “Have you created a flow chart or writing road map to help you organize your ideas?” The feedback also directed students toward relevant lessons or practice games. Importantly, feedback scaffolding helped to deliver only the most appropriate help; feedback was delivered only for the lowest threshold failed in the series of checks. We assumed that students who struggled to produce *any* text may not be ready to implement feedback about cohesion. Instead, these students may gain more from planning. If essays passed basic thresholds, they received feedback encouraging overall revision. Depending on the quality of individual sections, essays also received formative feedback for introduction,

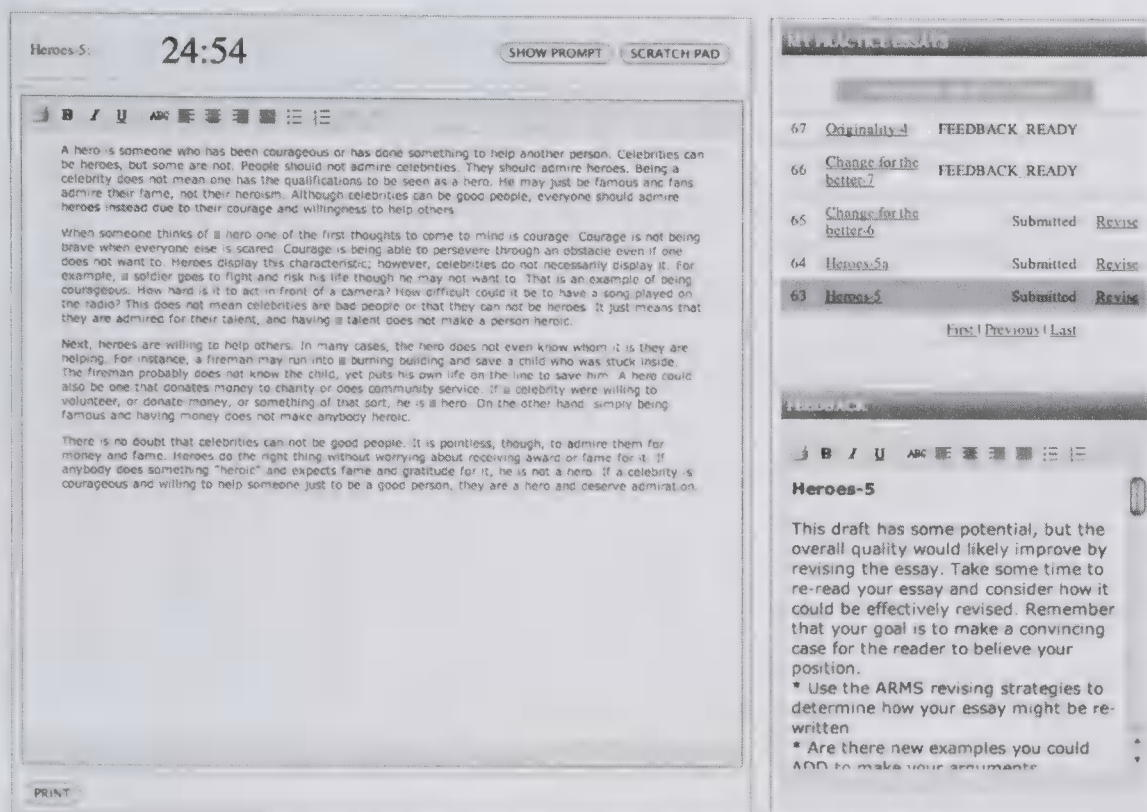


Figure 4. Essay Writing Interface.

body, and/or conclusion building strategies. For instance, an essay lacking a clear body might receive feedback stating “good writers often review their writing flowchart or an outline. Think about the best order and organization of the body paragraphs,” and asking, “Could a stranger understand your ideas without further explanation?” (Figure 5).

Unlike previous AWE systems, W-Pal focuses on strategy instruction and formative feedback and provides no specific error feedback on style, mechanics, spelling, or grammar. Spelling and grammar errors are relatively easy to detect, but assessing the quality and relevance of thesis statements, topic sentences, examples, counterarguments, and many other essay elements is more difficult. In the case of thesis statements, for example, it is a nontrivial matter to determine which sentence writers intended to communicate their position, if any. Once this determination is made, one must assess how the thesis relates to the prompt, subsequent arguments, and argument structure. At this stage of W-Pal development, we focused on the broader categories, which necessarily limited the specificity of W-Pal feedback.

In sum, development of W-Pal has sought to satisfy four central design principles that emerge from the ill-defined nature of writing, which has not been demonstrated in previous technologies for writing instruction. A fundamental question for deployment was whether an intelligent tutor for writing could be feasibly implemented with our target population of high school students. Would students use the system? Would students perceive a “computer tutor” as a viable instructional resource? To address these questions, we conducted a feasibility study in five high school English classrooms throughout a school year. Because our primary purpose was to assess feasibility, we did not employ a controlled experimental design (i.e., comparison to non-W-Pal instruction) or ablation design (i.e., selective removal of system features). Thus,

strong conclusions about efficacy cannot be drawn about the impact of W-Pal from this study.

Method

Participants

The intended users of Writing Pal are English-speaking high school students. Two high school English teachers and 141 10th grade students participated in this study over 6 months (November, 2010 to May, 2011) with their English classrooms. Teachers were asked to use the entire W-Pal, including Writing Strategy Modules, practice games, and essays. However, they were not given strict rules for how W-Pal was to be integrated (e.g., module order, assignment pacing and duration, or curriculum integration). Teachers and students (via their teachers) could contact the W-Pal team for technical support and teachers had weekly conference calls with the researchers. The participating high school was located in the Washington, DC area, and enrolled over 2,400 students. The school enrolled 49.0% female students, with 22.3% Asian, 4.2% Black, 9.0% Hispanic, and 59.9% White students; 7.0% of students were described as limited English proficiency, and 10.9% qualified for free or reduced-price meals.

Measures

Data logging. As students interacted with W-Pal, their access of system tools was logged. To examine usage of W-Pal, we considered access and completion of the lesson videos, frequency of games played, and frequency of essay submissions.

Lesson perception survey. After viewing each lesson, a five-item survey appeared. Using 4-point scales, students rated “how

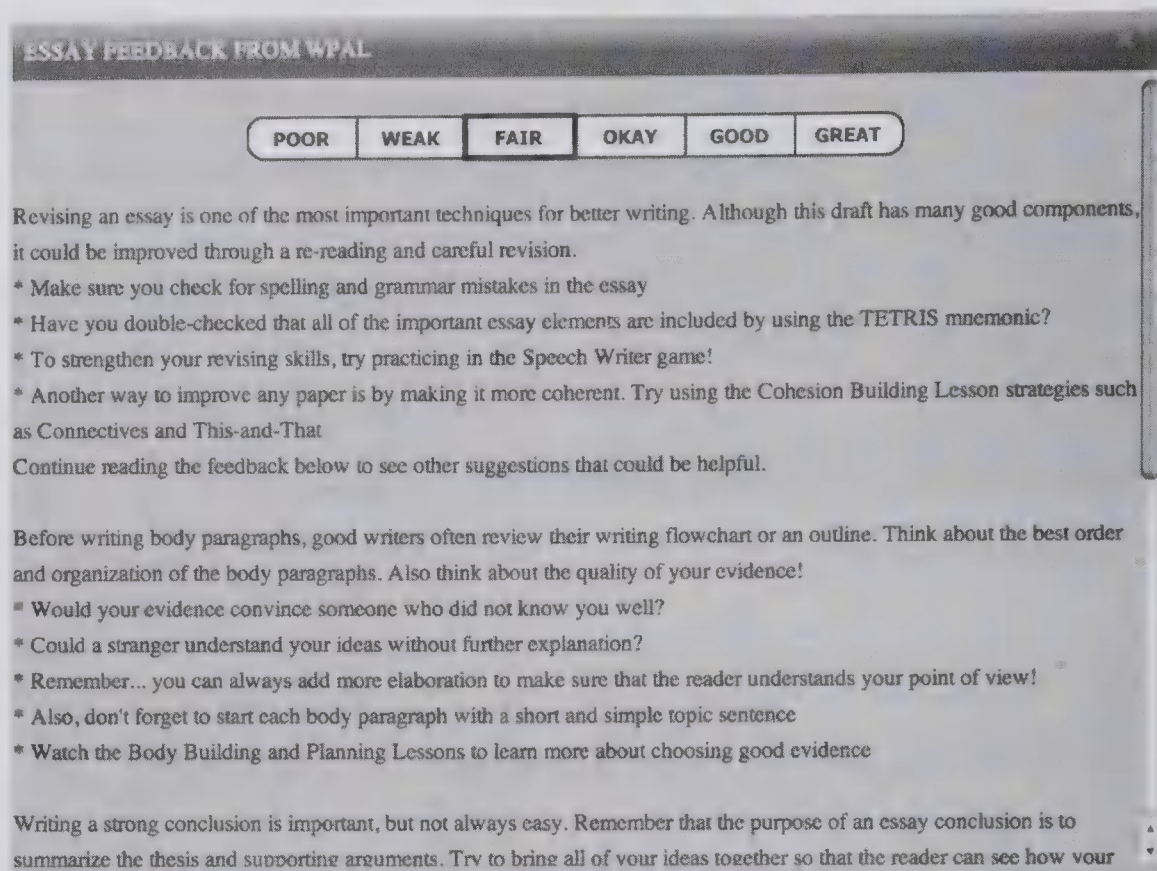


Figure 5. Example essay feedback report. WPAL = Writing Pal.

many new ideas" they learned (i.e., 0, 1–2, 3–4, or 5 or more ideas) and whether they would be willing view the lesson again. In open-ended items, students were asked to describe the "most helpful information" they learned, describe their perceptions of the animated characters, and provide suggestions for "how to improve this lesson."

Game perception survey. After interacting with W-Pal for several months (i.e., in February), students were asked to complete a four-item feedback survey of their perceptions of the games. Using 4-point scales, students rated a sampling of 11 games regarding helpfulness for practicing writing strategies, and rated the games regarding enjoyment. In two open-ended items, students were asked to provide suggestions for improving the helpfulness of the games and redesigning the games to be more enjoyable and engaging.

Feedback perception survey. In addition to the Game Perception Survey, students completed an eight-item survey of their perceptions of the essay writing tools and feedback. Using 4-point scales, students rated the overall difficulty of using the essay writing interface, the difficulty of specific tools, feedback quantity, understandability of the feedback, and usability of the feedback. In two open-ended items, students were asked to offer suggestions for making the feedback "more clear, more understandable, or more usable" and to suggest what "essay features or writing strategies" should be included in future feedback.

Pre- and post-study essays. Students wrote timed (25 min), prompt-based essays on two SAT-style prompts regarding "competition" and the influence of "images and impressions." These essays were written offline (i.e., not within W-Pal), manually transcribed by the research team, and scored via

natural language algorithms powered by Coh-Metrix (Crossley, Roscoe, Graesser, & McNamara, 2011). The accuracy of this algorithm, based on a separate test set of 105 essays and expert human scores, was 39% perfect agreement and 92% adjacent agreement. Descriptive information was also calculated for each essay, including the number of words, sentences, paragraphs, and sentences per paragraph. Text cohesion was assessed in terms of argument overlap (i.e., average overlap between head nouns and pronouns in adjacent sentences), given/new information (i.e., a Latent Semantic Analysis score indicating the amount of given compared to new information), and lexical diversity (i.e., degree to which a variety of words versus the same words are used across the text, using the measure D; Malvern, Richards, Chipere, & Durán, 2004). Prior research has indicated that higher quality essays are associated with a decrease in cohesion and an increase in lexical diversity (Crossley & McNamara, 2011). We also examined measures of lexical sophistication typically associated with essay quality (e.g., Crossley, Weston, McLain Sullivan, & McNamara, 2011), including word concreteness, word hypernymy (i.e., specificity), and the number of hedging words (i.e., an indicator of uncertainty).

Procedures

Students wrote a pre-study essay in November. Throughout the school year, teachers incorporated W-Pal into their English classroom curriculum. Students viewed the lessons, played the games, wrote practice essays, wrote essays assigned by teachers, and completed the surveys. Essays assigned by the teachers

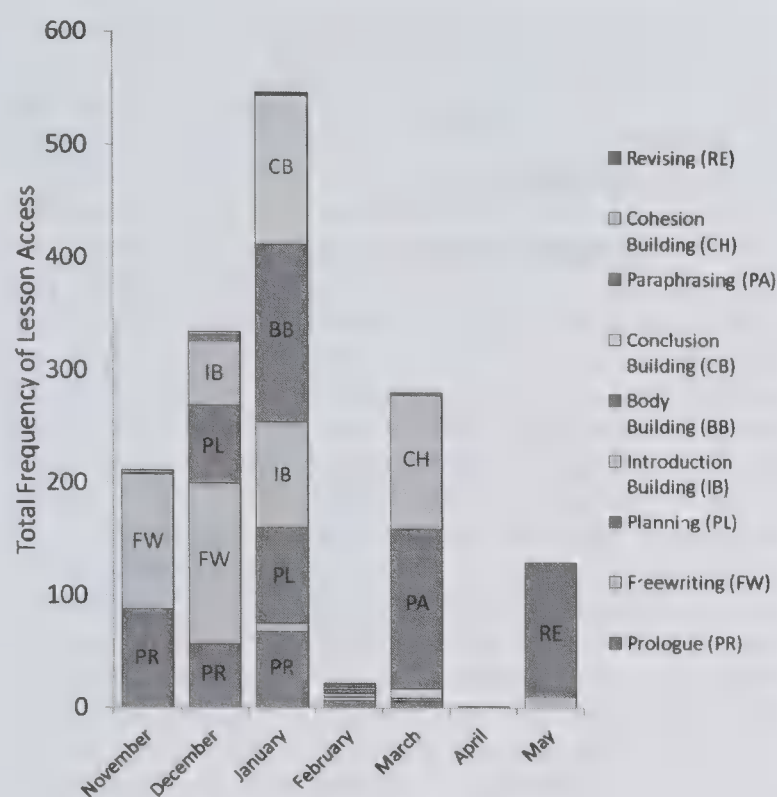


Figure 6. Total frequency of lesson viewing across a 6-month time period.

often explicitly linked to reading assignments, such as Moliere's *Tartuffe*. Students wrote a post-study essay in June. As this was an ecological setting, some students did not complete all assignments.

Table 2

Average Completion Percentage for Lesson Videos, Frequency of Game Play, and Maximum Number of Game Plays by Module

Module and game	Lesson completion		Game play		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	Maximum
Prologue	86.0	27.5			
Freewriting	90.2	28.7			
Freewrite Feud			0.59	0.89	4
Freewrite Fill-In			0.59	0.85	3
Planning	83.4	35.9			
Mastermind Outline			0.72	0.98	6
Planning Pump			0.76	0.82	4
Introduction Building	82.9	31.7			
Dungeon Escape			0.86	0.88	4
Essay Launcher			0.45	0.60	4
Fix-It			0.49	0.61	3
Body Building	82.9	37.2			
RAM-5			0.15	0.36	1
Fix-It			0.29	0.45	1
Conclusion Building	73.1	44.1			
Dungeon Escape			0.53	0.77	4
Fix-It			0.36	0.51	2
Paraphrasing	78.2	40.7			
Adventurer's Loot			0.44	0.51	2
Map Conquest			0.53	0.68	5
Cohesion Building	54.3	49.3			
CON-Artist			0.31	0.56	4
Undefined & Mined			0.54	1.14	6
Revising	68.5	45.1			
Speech Writer			0.29	0.54	3

Results

Students' Use of the System

Students interacted with W-Pal for about 16 total hours, on average, but students' use of W-Pal was unevenly distributed by module and across time. Figure 6 shows the distribution of strategy lessons accessed over the 6 months of the study (substantive activities are labeled with an abbreviation of the module name). Access was defined as a student interacting with at least one complete segment of the lesson. One pattern is that teachers mainly followed the sequence of prewriting, drafting, and revising. That is, they assigned the modules linearly in the "order" they were listed in the W-Pal interface. Teacher interviews indicated that they discouraged exploration; they preferred students to focus on current assignments and not to "get ahead." Second, most use of W-Pal lessons occurred during the first 3 months and then became more sporadic. January was particularly active as teachers encouraged students to complete the prewriting and drafting modules in preparation for SAT practice tests. Teachers did not assign lessons during February and April. Teacher interviews indicated that these months were devoted to separate writing assignments (e.g., a "how-to" paper), literature instruction (e.g., *Tale of Two Cities* and *Things Fall Apart*), and preparation for state exams.

Over time, lesson activity appeared to decrease. This pattern is substantiated by the average completion percentage of each module (Table 2). In general, students seemed more likely to complete the earlier modules (e.g., Freewriting), but tapered off in the later modules (e.g., Revising). One explanation may be student fatigue. After 5 months of using W-Pal, any novelty had likely diminished. In addition, teachers' focus on literature assignments and test

preparation may have led to a decreased emphasis of W-Pal in the classroom.

Figure 7 provides a similar visualization of students' game playing across modules, with substantive activity labeled by module. Few games were played in November, as most students had not unlocked any games. However, more games were played in the following months once teachers assigned the planning and drafting modules. Interestingly, game play continued during February when no new modules were assigned. Interviews revealed that teachers encouraged students to use the games as further practice during this time. In the final months, however, students mainly accessed the games associated with assigned modules. Table 2 shows the mean frequency of playing each game. Games encountered earlier in instruction (e.g., Mastermind Outline), were played slightly more often than later games (e.g., Speech Writer). However, there was variation in game play and some games from later modules were played as often as earlier games. The overall low frequency of play is likely a result of teachers' discouragement of exploration. The wide variety of games offered by W-Pal may have also contributed. With many games to choose from, the desire to "master" any one game might have been low.

Use of the essay writing tools was somewhat sparse because teachers used W-Pal for specific assignments rather than self-selected practice. Teachers assigned two to three W-Pal practice essays with automated feedback (on "Honesty," "Uniformity," or "Heroes") in December and January. Students were not required to revise these essays and course grades were based only on assignment completion. In April and May, teachers assigned students to write on the "Memories" prompt in relation to the novel *Things Fall Apart* (with automated feedback). Revising of this essay occurred outside of W-Pal via extensive peer reviewing. Teachers initially reported confusion about how teacher-created prompts differed from built-in W-Pal writing prompts—essays written on teacher-created prompts could not be assessed by the algorithm in this version of the system. However, after discussion about this

functionality, teachers still chose to create two new assignments in W-Pal. In one essay, students wrote about interpersonal perceptions in relation to the novel *Tartuffe* (January), and students responded to a newspaper article about the value of study halls in high schools (February).

Interviews revealed that teachers perceived W-Pal's essay tools favorably, and felt that the system allowed them to assign more writing that was feasible without W-Pal. Specifically, W-Pal provided an accessible means for students to practice writing, with automated feedback, and teachers could access these essays and feedback online. W-Pal also provided several ready-made writing prompts for assignments. However, the system could not support the full range of writing assignments that were required in the curriculum, a common problem for AWE systems (e.g., Grimes & Warschauer, 2008). Different writing genres (e.g., journalism and narrative) possess unique constraints that cannot be assessed by the same algorithm; computational linguistics models must tailored to each type. Most systems, including W-Pal, have focused upon persuasive writing due to its importance for standardized testing. Other genres are not currently supported but are a target for future development. Teachers also understood that W-Pal was still "in development" and thus were somewhat wary of basing students' grades on W-Pal assessments. This concern may also have limited the number of practice essays teachers assigned. Teachers may have been hesitant to utilize W-Pal for writing practice unless they could also review or grade the assignments independently. Teachers understood the scoring and feedback procedures but, as conscientious instructors, they wanted to remain actively aware of and involved with their students' work and progress.

In sum, students used a variety of W-Pal features but did so unevenly over the year. W-Pal deployment was not a smooth and continuous process; as with any educational resource, teachers were selective and opportunistic about how and when to use the system. Results also suggest that engagement with the system declined over time. We next consider students' perceptions of W-Pal and how such perceptions may have impacted system use and feasibility.

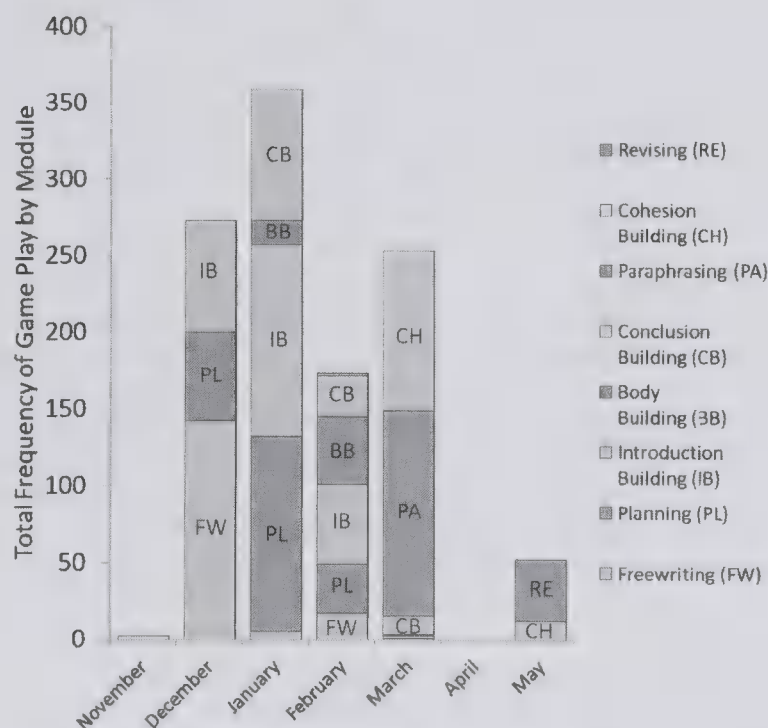


Figure 7. Total frequency of game playing across a 6-month time period.

Lesson Perceptions

Figure 8 (left side) presents the percentage of students as a function of the number of ideas they reported having learned from the lessons. In general, students reported the lessons to be helpful and informative. On average and across lessons, over half of the students (55.8%) reported learning three or more ideas per lesson. Within the open-ended questions asking students to summarize the *most helpful idea* learned from the lessons, the mnemonic devices were the most frequent response. Thus, students seemed to value and remember the acronyms such as *TAG*, *RECAP*, and *ARMS* designed to cue recall of specific strategies. In contrast, students disliked the presentation of the lessons (Figure 8, right side). On average and across lessons, many students viewed the characters as *awkward* (62.3%) and *boring* (60.6%), but still *informative* (30.4%).

In open-ended responses (see Table 3), students critiqued agent dialog and requested succinct instruction with more competent and less "cartoonish" characters. The computerized voices were also unpopular, in part because of a text-to-speech glitch that sometimes caused overlapping speech. Both students and teachers re-

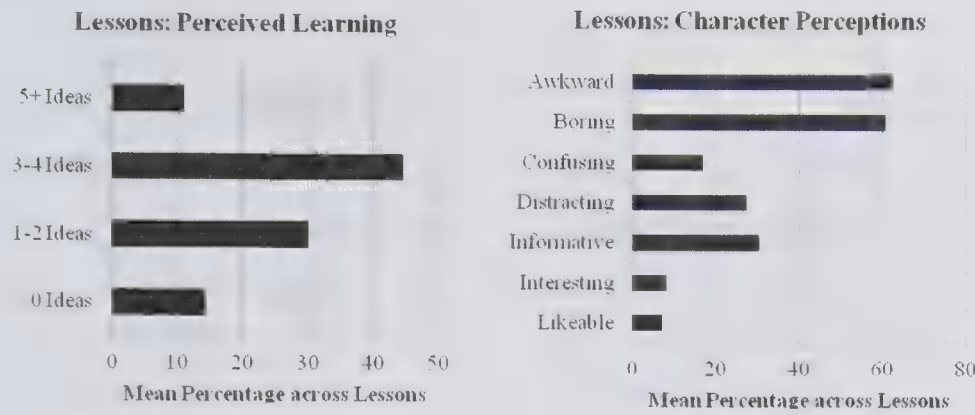


Figure 8. Student perceptions of learning and animated characters in Writing Pal lessons.

quested that the lessons be shorter and faster, while retaining all of the information. These concerns are summarized by one student, who commented, “the little jokes between characters aren’t amusing, especially in the monotone computer voices. Cutting out all the unnecessary dialogue between characters would shave off a good amount of time.” Altogether, these results support the earlier hypothesis that lesson use decreased due to fatigue. As students progressed through the lessons, and encountered the same design issues, students’ willingness to engage with the lessons likely decreased.

Game Perceptions

Across sampled games, students ($n = 116$) reported the games to be *somewhat helpful* (50.5%) or *very helpful* (29.6%)

for practicing the writing strategies (Figure 9, left). Similarly, students reported the games to be *somewhat enjoyable* (46.4%) or *very enjoyable* (19.1%) to play (Figure 9, right). Thus, most students felt that the games they played were beneficial and generally engaging. Open-ended comments (see Table 3) highlighted ways in which students felt the games could be improved. For example, one student requested that we make the games “*more challenging* [because even] if I hadn’t taken the W-Pal lessons, I would have been able to complete the challenges with fairly high scores.” Other students expressed interest in further generative practice, such as “when we learn the strategies, I think should be a challenge where we actually use the strategy instead of finding them in essays.” Another student suggested that “the games could be more difficult and more

Table 3
Student Responses and Recommendations Regarding Strategy Lessons and Practice Games

Observation	Examples
1. Students valued the strategies and mnemonics.	<p>“FAST PACE is going to help me write better essays! I learned important acronyms, and information. I learned to think about the prompt, add questions, think about the opposing side”</p> <p>“The TAG mnemonic and the attention grabbing techniques were very helpful for making me understand introductions better”</p> <p>“RECAP—restate, explain ideas, closing, avoid new things, present interestingly”</p>
2. Students disliked the length and presentation style of the lessons.	<p>“Their voices are very robotic and the lesson was way too long, maybe if it was split into several sections then it would be easier to concentrate on the task”</p> <p>“Had very good information but I disliked the synthesized voices”</p> <p>“The information is good but I lost interest throughout the lesson. I feel like I would learn a lot more if the information went faster and was straightforward”</p>
3. Students desired games that were more difficult and interactive.	<p>“When we learned the strategies, I think there should be a challenge where we actually use the strategy instead of finding them in essays”</p> <p>“Make the challenges more challenging. Even if I hadn’t taken the W-Pal lessons, I would have been able to complete the challenges with fairly high scores”</p>
4. Some students found the game instructions inadequate.	<p>“Some of the instructions were hard to follow”</p> <p>“I had a little trouble understand exactly what to do with the directions.”</p>
5. Students suggested improvements in the game graphics and sound.	<p>“The games could have better graphics and music to make the games more enjoyable”</p> <p>“The games are slow and the graphics are not the best, so unfortunately, the games become boring which weakens their effectiveness”</p>
6. Students requested that more game elements be added.	<p>“Many of the games were not very fun because they had a learning element that was very obvious. It would be better if the element was not as obvious, so the game was more fun. Basically, more pictures and music and less words”</p> <p>“Make it a point system and make it a competition amongst our peers”</p>

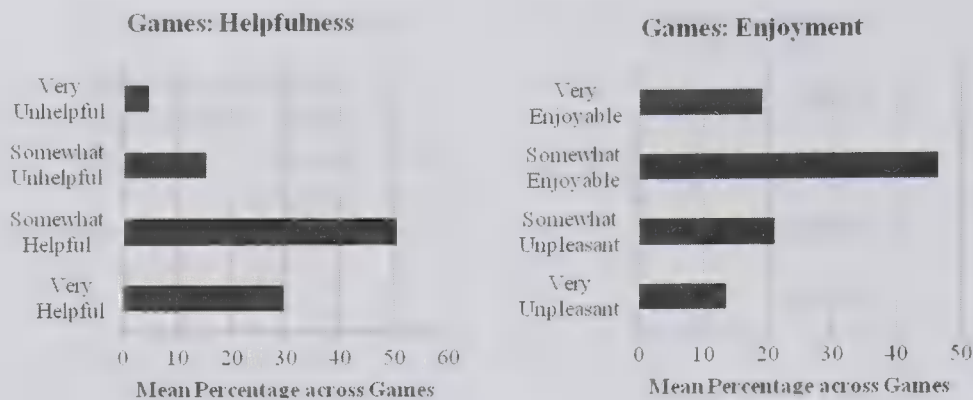


Figure 9. Student perceptions of the helpfulness and enjoyment of games.

interactive for learning these writing strategies rather than just reading.” In sum, students valued the games, but positive perceptions may have been impacted by games that lacked challenge, opportunities for interaction, or clear directions.

Essay Writing and Feedback Perceptions

Overall, students ($n = 103$) rated the essay writing tools as *easy* or *very easy* (81.5%) to use (Figure 10, top left). However, two features frustrated some students: 23.7% of students reported that reading the feedback was *somewhat* or *very difficult*, and 24.6% felt that revising their essays was *somewhat* or *very difficult*. This may have been due to feedback quantity or clarity (Figure 10, top right). Although most students reported that they received *just the right amount of feedback* (49.5%), others reported that they received *not enough* (38.8%) or *too much*

(11.6%). From internal testing (Roscoe, Varner, Cai, Weston, Crossley, & McNamara, 2011), we knew that feedback quantity could be variable. Essays that failed a basic check (e.g., length) received only one feedback message. However, essays that advanced further could receive more messages on multiple topics. These extremes may have led to perceptions of insufficient or overwhelming feedback, respectively. Similarly, as shown in Figure 10 (bottom left), most students rated the feedback as *understandable* (61.2%), but some students rated the feedback as *somewhat confusing* (29.1%) or *very confusing* (9.71%). Despite these challenges, students rated the feedback as useful (Figure 10, bottom right) *occasionally* (45.6%) or *often* (33.0%).

Students’ open-ended responses (see Table 4) further highlighted student concerns. Specificity was a particular critique;

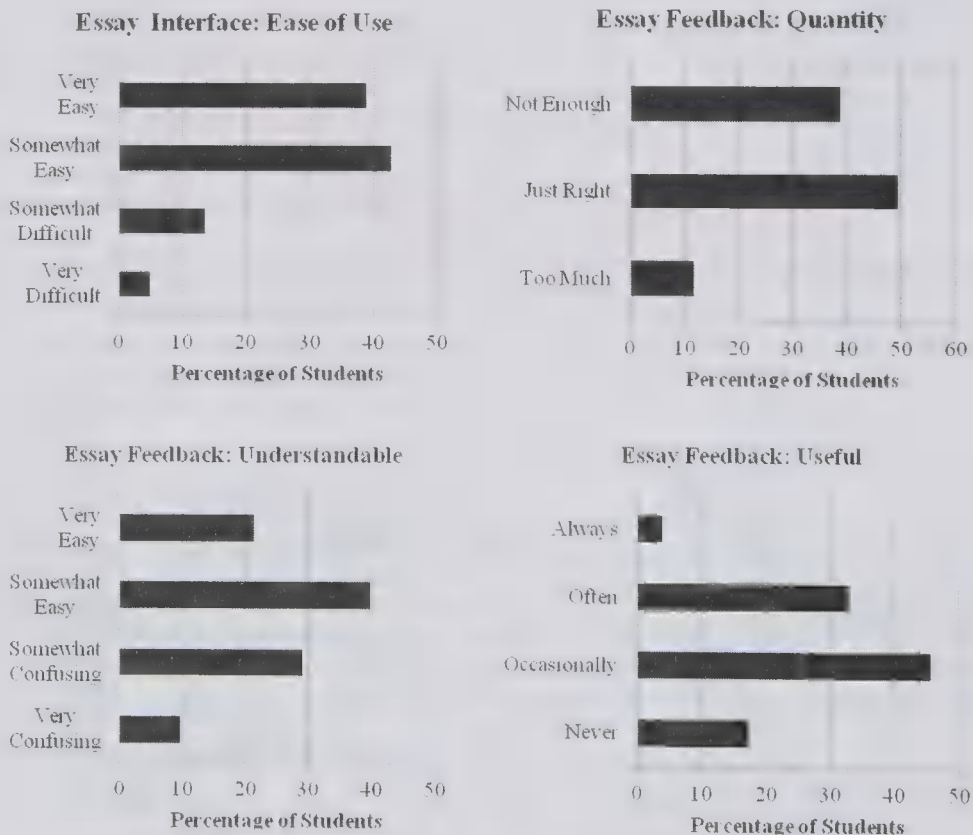


Figure 10. Student perceptions of ease of use, quantity, understandability, and usefulness of automated essay feedback.

Table 4
Student Responses and Recommendations Regarding Essay Scoring and Feedback

Observation	Examples
7. Students requested more specific feedback.	"The feedback should show what specific things made me get the grade" "The feedback needs to be more helpful for us on our own personal essay. Not just general feedback. I don't know what I did wrong in my essay when you just give a general understanding of it"
8. Students requested more individualized feedback.	"My introduction and my supports. I still have a hard time finding supports that directly answer the question" "I would like to know in the feedback if my examples were not strong enough, if I had a weak thesis, things like that"
9. Students expressed conflicting concerns about the quantity of feedback.	"Use less feedback and cut straight to the point of what the essay needs and give examples" "The feedback is very brief. W-Pal never really tells you what you need to improve on."
10. Some students expressed skepticism at the speed and accuracy of scoring.	"I do not like how the essay is graded in less than a second! I feel my essay is not being graded properly and I don't feel I have been given accurate feedback" "You cannot grade an essay in 5 seconds! Everybody gets the same grading of "fair." I can't use it if I don't believe that it is true."

students commented that "the constructive criticism could be a bit more detailed on what the writer needs to work on instead of an overview" or should "be specific and give us exact examples on what we should do to improve our writing." Other students requested that the feedback system provide information on both the strengths and weaknesses in an essay, e.g., "the automatic feedback could also give you good points on your essay, what was strong and what you should continue do." Thus, the provision of feedback at the level of broad categories (e.g., body building strategies) rather than specific essay elements (e.g., evidence quality) was helpful but inadequate for some students. Overall, the feedback provided by W-Pal in this study was perceived as beneficial and relevant to students' needs, but the content of the feedback should be expanded to address more detailed issues.

Essay Quality

The natural language algorithm analyses of pre-study and post-study essays ($n = 113$) are provided in Table 5. Essay scores

increased significantly from a mean of 2.3 ($SD = 0.8$) prior to the study to a mean of 2.9 ($SD = 0.8$) after the study, $t(112) = 5.85$, $p < .001$, $d = 0.71$. Associated with these gains were positive changes in essay structure and lexical sophistication (see Table 5). Post-study essays were longer, containing more words and sentences. Essays also showed a clearer paragraph structure, with more paragraphs overall and somewhat fewer sentences per paragraph (e.g., fewer students wrote one-paragraph essays). Post-study essays improved in vocabulary use, including more concrete wording, more precise wording (word hypernymy), fewer hedging words (e.g., *maybe* or *might*), and greater diversity. Finally, essays showed more developed and elaborated content with less repetition of themes (less overlap of arguments and given information) and wording (increased lexical diversity).

Given the patterns of W-Pal use throughout the feasibility study, it would be unlikely to observe strong effects of using the system on essay gains. W-Pal was only one component of a broader curriculum. Nonetheless, to assess how and whether use of W-Pal

Table 5
Essay Characteristics for Pre- and Post-Study Timed Essays

Measure	<i>M (SD)</i>		<i>t</i> (112)	<i>p</i>
	Pre	Post		
Essay score	2.30 (0.84)	2.88 (0.79)	5.85	<.001
Length				
Number of words	260.81 (76.38)	308.27 (84.49)	6.49	<.001
Number of sentences	15.46 (5.10)	18.27 (5.13)	5.66	<.001
Structure				
Number of paragraphs	3.43 (1.32)	3.97 (0.83)	3.87	<.001
Sentences per paragraph	5.33 (3.02)	4.72 (1.44)	-1.83	.071
Cohesion ^a				
Argument overlap	0.51 (0.17)	0.41 (0.14)	-5.04	<.001
Given/new information	0.32 (0.04)	0.30 (0.04)	-4.43	<.001
Lexical diversity	85.13 (21.39)	98.29 (21.56)	5.27	<.001
Lexical sophistication				
Word concreteness	387.00 (32.29)	405.53 (30.78)	3.87	<.001
Word hypernymy	1.57 (0.23)	1.66 (0.19)	4.23	<.001
Hedging words	14.2 (10.6)	9.9 (7.6)	-4.10	<.001

^aThese cohesion indices indicate the extent to which arguments, ideas, and words are repeated across sentences and throughout the text.

might have influenced writing proficiency, an exploratory linear regression analysis was conducted to identify potential predictors of post-study essay quality. Eight predictor variables were simultaneously entered. As measures of students' prior writing ability and knowledge, *pre-study essay scores* and self-reported *grade-point average* (GPA) were included. As indicators of system use, we included students' percentage completion of *prewriting lessons* (Freewriting and Planning), *drafting lessons* (Introduction Building, Body Building, and Conclusion Building), and *revising lessons* (Paraphrasing, Cohesion Building, and overall Revising). Similarly, we included the frequency of game play within each phase: *prewriting games* (Freewrite Feud, Freewrite Fill-In, Mastermind Outline, and Planning Pump), *drafting games* (Essay Launcher, Dungeon Escape, Fix It, and RAM-5), and *revising games* (Adventurer's Loot, Map Conquest, Undefined & Mined, CON-Artist, and Speech Writer). Because teachers chose to restrict essay writing practice, there was little variability in essay writing, and this variable was not included.

The resulting linear regression model was significant, $F(112) = 2.93$, $p = .005$, $R^2 = .18$, accounting for about one fifth of the variance in post-study essay scores (see Table 6). Two variables were predictive of essay quality: *pre-study essay scores* and viewing of the *drafting lessons*. Interestingly, students' prior writing ability (pre-study essay score), but not their GPA, was a significant predictor of post-study essay quality. These results suggest that writing skill was not solely a function of students' prior academic abilities, but reflected knowledge of specialized skills and strategies related to writing. Students' completion of the drafting lessons was positively associated with their writing development above and beyond prior writing ability. Drafting lessons are perhaps the most immediately relevant to students' writing of timed essays, because they provide direct strategies for generating essay text. Overall, although we cannot conclude that W-Pal directly improved students' writing, these results tentatively support the feasibility of intelligent tutoring of writing in high school classrooms.

Discussion

The unique design of W-Pal was informed by the ill-defined nature of writing, in which there is significant ambiguity and subjectivity with respect to pedagogy and assessment. We have sought to provide comprehensive and modular strategy instruction, diverse opportunities for extended practice, and formative feedback on students' writing. In this study, we evaluated how W-Pal was perceived by high school in English classrooms. A fundamen-

tal assumption was that feasibility depends on whether users view the system as a valid and valuable tool for instruction and feedback. Thus, students' use and perceptions of W-Pal were the central focus.

Our results suggest that this initial version of W-Pal was generally well received. Most components of W-Pal were judged as beneficial sources of writing instruction, practice, and feedback. Students could describe specific content that they learned from the lessons and games, and rated these tools and essay feedback as helpful and easy to use. Students seemed to view a "computer tutor" as a worthwhile addition to the English classroom curriculum. Preliminary evidence also suggests that students benefitted from using certain W-Pal tools. Thus, the initial iteration of W-Pal was feasible with regards to positive user perceptions and usage.

Our results also highlighted several problems to overcome that may undermine long-term feasibility and potential efficacy. First, students felt that the lessons were too long and didactic, and disliked the cartoonish characters in the lessons. In some ways, the lengthy lesson videos were too similar to a *presentational* mode of writing instruction described by Hillocks (1984). Hillocks contrasted writing outcomes for interventions that employed different instructional modes and content. The most effective instruction occurred in an *environmental* mode wherein instructors minimized lecturing and focused on specific objectives and strategies, with ample opportunities for scaffolded practice. In contrast, instruction was less effective in the prescriptive and teacher-dominated *presentational* mode. Although interactive checkpoints were included in the lessons, students' overall perceptions were that the lessons were too long, boring, and lecture-like. This lesson structure may also have insufficiently met the goal of providing modular instruction; each lesson video comprised multiple strategies related to multiple goals. A series of shorter lessons, each with a focus on one or two related strategies, may have been more germane to Hillocks' *environmental* mode. Students could iterate between lessons and practice more flexibly, and instructors could be more selective with the content they wished to cover.

More broadly, the issue of information density within instructional modules speaks to the appropriate grain-size of ITS instruction in ill-defined domains. When learners must make many strategic decisions to enact a task, instruction may need to focus initially on fewer decisions before asking students to synthesize them. With each additional, simultaneous strategy choice, it becomes more difficult for learners to perceive the impact or utility of each strategy. In problem-solving domains (e.g., physics), re-

Table 6
Linear Regression Analysis to Predict Post-Study Essay Scores

Variable	<i>r</i>	<i>B</i>	β	<i>SE</i>	<i>t</i>	<i>p</i>
Pre-study essay score	.31	0.273	.293	.088	3.09	.003
GPA	.12	-0.031	-.026	.118	-0.26	.794
Prewriting lessons	.05	-0.002	-.131	.002	-1.02	.308
Drafting lessons	.17	0.004	.431	.001	2.81	.006
Revising lessons	.08	-0.001	-.157	.001	-1.20	.233
Prewriting games	-.07	-0.015	-.053	.032	-0.47	.640
Drafting games	.00	-0.054	-.191	.035	-1.54	.126
Revising games	.12	0.058	.181	.037	1.59	.115

Note. GPA = grade-point average. Estimated constant term is 2.32. Boldface font indicates statistically significant predictors.

search has shown benefits for systems that require students to specify each step of their solution process rather than merely the final answer (Hausmann, VanLehn, Nokes, & Gershman, 2009; VanLehn et al., 2005). This decomposition allows the system to assess and provide feedback for individual steps, and learners are encouraged to consider the impact of each decision. Analogously, in intelligent tutoring in ill-defined domains such as writing, it may be beneficial to teach fewer writing strategies at one time so that students can more gradually build up to the full complexity of the writing process. The modular content of the ITS should facilitate the decomposition of complex processes into manageable units for initial learning, which can be subsequently recombined and applied strategically in later practice.

A second critique expressed by students related to types and difficulty of learning tasks presented in the educational games. Surprisingly, some students expressed interest in more difficult games that required active generation of text. Such students wanted to practice by applying strategies to their own writing rather than inspecting examples written by others. Not surprisingly, we also observed a high degree of variability in students' game preferences. Games that were played frequently or rated highly by some students were despised by others, and vice versa. Only a few games were broadly disliked; for instance, *RAM-5* (a body building game in which students matched potential evidence to topic sentences) had little replay value, and the task was vague. A few games were liked by the majority of students. One example was *Map Conquest*, a Risk-like game in which students earn resources by identifying paraphrasing strategies and then use those resources to "conquer" a map controlled by computer opponents. An interesting facet of this game is that the learning task (identifying paraphrases) and the game task (taking over the map) are disjoint. Success in the learning task did not guarantee success in the game, and vice versa. This might have made the "gaming" aspects of the practice more salient for some students.

The positive perception of educational games in W-Pal suggests that this could be a valuable component for intelligent tutoring in ill-defined domains. Specifically, games may help to offset some of the motivational threats that undermine students' engagement with ITSs and extended practice. Success in ill-defined domains requires learning of underspecified concepts and relations, and the ability to recharacterize problems to apply available strategies (Lynch et al., 2009). Developing such skills may be frustrating as students struggle to master many decisions and tasks. Indeed, students often report high apprehension and low confidence regarding their writing abilities (e.g., Pajares, 2003). Our results hint that educational games may help to ameliorate some of the affective challenges that arise with learning in ITSs and ill-defined domains (e.g., Craig, Graesser, Sullins, & Gholson, 2004). Games may provide a more pleasant setting where practice is embedded within an enjoyable experience, and feedback is framed within game mechanics or narrative rather than overt critique. However, based on these findings, developers who wish to bolster ITSs with educational games should ensure that the games offer sufficient challenge, promote generative activity, and exhibit varied gameplay.

A final concern revealed by the study, and perhaps the greatest challenge for future development, was the need for more specific and individualized feedback. Students expressed a clear desire to learn more about the individual strengths and weaknesses of their

essays, and a lack of such specificity undermined confidence in the system for some students. However, improvements to W-Pal's feedback engine will require sophisticated additions and refinements to underlying computational linguistics algorithms. Although the framing and content of the feedback is paramount—feedback must be well-constructed to provide actionable suggestions in a scaffolded and nonthreatening manner—the feedback process is necessarily constrained to essay features that can be reliably detected. We are currently exploring alternative methods for developing feedback algorithms.

Issues of valid and formative feedback generalize beyond essays and writing. Algorithm development is likely to be a key obstacle in the growth of tutors for writing and other ill-defined domains (McNamara et al., in press). Any ITS that accepts open-ended or natural language input, and attempts to respond to learners with intelligent guidance and help, may need to solve a similar set of problems. For example, an ITS that allows users to explain scientific concepts will require algorithms that can process and interpret users' intended answers. Tutorial feedback, such as corrective hints or explanations, will be more valuable to the extent that users believe the system can target their individual strengths, weaknesses, knowledge, and misconceptions.

Conclusion

W-Pal development and testing have revealed several issues and lessons for building an ITS in ill-defined domains. Some of these feasibility problems may be termed *presentational*, in that they can be overcome by redesigning the interface or mode of instruction to be more modular, engaging, succinct, game-like, and so on. These are relatively easy to fix—more recent iterations of W-Pal have already addressed a number of concerns—although they are often only revealed through extensive usability and feasibility testing. Other feasibility issues may be termed *algorithmic* and relate to the methods by which complex, open-ended, and ambiguous student inputs are processed and evaluated. New and innovative methods for assessing such inputs may be required to realize the full potential of intelligent tutoring in ill-defined domains. However, in ill-defined domains, a certain level of permanent ambiguity may have to be embraced, and the focus must be on guiding students toward progress and independence, rather than delivering, correcting, or testing a well-defined body of knowledge.

References

- Aleven, V., & Koedinger, K. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science*, 26, 147–179. doi:10.1207/s15516709cog2602_1
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V. 2. *Journal of Technology, Learning, and Assessment*, 4(3). Retrieved from <http://www.jtla.org>
- Beal, C., Arroyo, I., Cohen, P., & Woolf, B. (2010). Evaluation of AnimalWatch: In intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning*, 9, 64–77.
- Bell, C., & McNamara, D. (2007). Integrating iSTART into a high school curriculum. *Proceedings of the 29th annual meeting of the Cognitive Science Society* (pp. 809–814). Austin, TX: Cognitive Science Society.
- Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing system. *AI Magazine*, 25, 27–36.

- Caccamise, D., Franzke, M., Eckhoff, A., Kintsch, E., & Kintsch, W. (2007). Guided practice in technology-based summary writing. In D. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 375–396). Mahwah, NJ: Erlbaum.
- Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29, 241–250. doi:10.1080/1358165042000283101
- Crossley, S., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21, 170–191. doi:10.1504/IJCEELL.2011.040197
- Crossley, S., Roscoe, R., Graesser, A., & McNamara, D. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. *Proceedings of the 15th international conference on artificial intelligence in education* (pp. 438–440). Auckland, New Zealand: AIED.
- Crossley, S., Weston, J., McLain Sullivan, S., & McNamara, D. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, 28, 282–311. doi:10.1177/0741088311410188
- Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill* (Research Report No. RR-08–55). Princeton, NJ: Educational Testing Service.
- De la Paz, S., & Graham, S. (2002). Explicitly teaching strategies, skills, and knowledge: Writing instruction in middle school classrooms. *Journal of Educational Psychology*, 94, 687–698. doi:10.1037/0022-0663.94.4.687
- Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1), Retrieved from <http://www.jtla.org>
- Flower, L., & Hayes, J. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32, 365–387. doi:10.2307/356600
- Gamper, J., & Knapp, J. (2002). A review of intelligent CALL systems. *Computer Assisted Language Learning*, 15, 329–342. doi:10.1076/call.15.4.329.8270
- Graesser, A., Lu, S., Jackson, G., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments & Computers*, 36, 180–192. doi:10.3758/BF03195563
- Graesser, A., & McNamara, D. (2012). Use of computers to analyze and score essays and open-ended verbal responses. In H. Cooper, P. Camic, R. Gonzalez, D. Long, & A. Panter (Eds.), *APA handbook of research methods in psychology* (pp. 307–325). Washington, DC: American Psychological Association.
- Graesser, A., McNamara, D., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point & Query, AutoTutor, and iSTART. *Educational Psychologist*, 40, 225–234. doi:10.1207/s15326985ep4004_4
- Graham, S., McKeown, D., Kiuhara, S., & Harris, K. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, 104, 879–896. doi:10.1037/a0029185
- Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99, 445–476. doi:10.1037/0022-0663.99.3.445
- Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment*, 8, 4–43.
- Hausmann, R., VanLehn, K., Nokes, T., & Gershman, S. (2009). *The design of self-explanation prompts: The fit hypothesis*. Paper presented at the 31st annual meeting of the Cognitive Sciences Society, Amsterdam, the Netherlands.
- Hillocks, G. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education*, 93, 133–170. doi:10.1086/443789
- Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, 47, 549–566. doi:10.2307/358601
- Jackson, G., & McNamara, D. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology*, 105, 1036–1049. doi:10.1037/a0032580
- Johnson, W., & Wu, S. (2008). Assessing aptitude for learning with a serious game for foreign language and culture. In B. Woolf, E. Aimeur, R. Nkambo, & S. Lajoie (Eds.), *Intelligent tutoring systems* (pp. 520–529). Berlin, Germany: Springer-Verlag. doi:10.1007/978-3-540-69132-7_55
- Kellogg, R., Whiteford, A., & Quinlan, T. (2010). Does automated feedback help students learn to write? *Journal of Educational Computing Research*, 42, 173–196. doi:10.2190/EC.42.2.c
- Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street®: Computer-guided summary writing. In T. K. Landauer, D. M. McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent semantic analysis* (pp. 263–277). Mahwah, NJ: Erlbaum.
- Landauer, T., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, 10, 295–308. doi:10.1080/0969594032000148154
- Landauer, T., Lochbaum, K., & Dooley, S. (2009). A new formative assessment technology for reading and writing. *Theory Into Practice*, 48, 44–52. doi:10.1080/00405840802577593
- Lynch, C., Ashley, K., Pinkwart, N., & Aleven, V. (2009). Concepts, structures, and goals: Redefining ill-definedness. *International Journal of Artificial Intelligence in Education*, 19, 253–266.
- Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke, England: Palgrave. doi:10.1057/9780230511804
- McGarrell, H., & Verbeem, J. (2007). Motivating revision of drafts through formative feedback. *ELT Journal*, 61, 228–236. doi:10.1093/elt/ccm030
- McNamara, D., Crossley, S., & McCarthy, P. (2010). Linguistic features of writing quality. *Written Communication*, 27, 57–86. doi:10.1177/0741088309351547
- McNamara, D., Crossley, S., & Roscoe, R. (2012). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-012-0258-1
- McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research*, 34, 147–171. doi:10.2190/1RU5-HDTJ-A5C8-JVWE
- McNamara, D., Raine, R., Roscoe, R., Crossley, S., Dai, J., Cai, Z., . . . Graesser, A. (2011). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298–311). Hershey, PA: IGI Global.
- Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability*. Retrieved from AQA Centre for Education Research and Policy website: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf
- Michael, J., Rovick, A., Glass, M., Zhou, Y., & Evens, M. (2003). Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments*, 11, 233–262. doi:10.1076/ilee.11.3.233.16543
- Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, 19, 139–158. doi:10.1080/10573560308222

- Proske, A., Narciss, S., & McNamara, D. (2012). Computer-based scaffolding to facilitate students' development of expertise in academic writing. *Journal of Research in Reading, 35*, 136–152. doi:10.1111/j.1467-9817.2010.01450.x
- Rock, J. (2007). *The impact of short-term use of Criterion on writing skills in 9th grade* (Research Report no. RR-07-07). Princeton, NJ: Educational Testing Service.
- Roscoe, R., Varner, L., Cai, Z., Weston, J., Crossley, S., & McNamara, D. (2011). Internal usability testing of automated essay feedback in an intelligent writing tutor. In R. Murray & P. McCarthy (Eds.), *Proceedings of the 24th international Florida Artificial Intelligence Research Society conference* (pp. 543–548). Menlo Park, CA: AAAI Press.
- Roscoe, R., Varner, L., Weston, J., Crossley, S., & McNamara, D. (in press). The Writing Pal Intelligent Tutoring System: Usability testing and development. *Computers and Composition*.
- Rowley, K., & Meyer, N. (2003). The effect of a computer tutor for writers on student writing achievement. *Journal of Educational Computing Research, 29*, 169–187. doi:10.2190/3WVD-BKEY-PK0D-TTR7
- Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of the Intelli-Metric essay scoring system. *Journal of Technology, Learning, and Assessment, 4*, 3–21.
- Shank, R., & Neeman, A. (2001). Motivation and failure in educational systems design. In K. Forbus & P. Feltovich (Eds.), *Smart machines in education* (pp. 37–69). Cambridge, MA: AAAI Press/MIT Press.
- Shermis, M., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.
- Shermis, M., Burstein, J., & Bliss, L. (2004). *The impact of automated essay scoring on high stakes writing assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Shute, V. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189. doi:10.3102/0034654307313795
- Simon, H. (1973). The structure of ill structured problems. *Artificial Intelligence, 4*, 181–201. doi:10.1016/0004-3702(73)90011-8
- VanLehn, K., Lynch, C., Schulze, K., Shapiro, J., Shelby, R., Taylor, L., . . . Wintersgill, M. (2005). The Andes Physics Tutoring System: Lessons learned. *International Journal of Artificial Intelligence in Education, 15*, 147–204.
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*, 157–180. doi:10.1191/1362168806lr190oa
- Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*, 22–36.
- Wolfe, C., Britt, M., Petrovic, M., Albrecht, M., & Kopp, K. (2009). The efficacy of a web-based counterargument tutor. *Behavior Research Methods, 41*, 691–698. doi:10.3758/BRM.41.3.691

Received December 15, 2011

Revision received December 18, 2012

Accepted February 11, 2013 ■

Correction to Hernandez et al. (2013)

In the article “Sustaining optimal motivation: A longitudinal analysis of interventions to broaden participation of underrepresented students in STEM” by Paul R. Hernandez, P. Wesley Schultz, Mica Estrada, Anna Woodcock, and Randie C. Chance (*Journal of Educational Psychology*, Vol. 105, No. 1, pp. 89–107 doi: 10.1037/a0029691), there was an error in the Appendix. The items listed below should have appeared without an asterisk.

TGO-6. An important reason I do my school work is because I enjoy it.

PAP-2. It’s important to me that the other students in my classes think that I am good at my work.

PAP-3. I want to do better than other students in my classes.

PAV-1. It’s very important to me that I don’t look stupid in my classes.

PAV-5. One reason I would not participate in class is to avoid looking stupid.

DOI: 10.1037/a0034254

Learning Intercultural Communication Skills With Virtual Humans: Feedback and Fidelity

H. Chad Lane, Matthew Jensen Hays, Mark G. Core, and Daniel Auerbach
University of Southern California

In the context of practicing intercultural communication skills, we investigated the role of fidelity in a game-based, virtual learning environment as well as the role of feedback delivered by an intelligent tutoring system. In 2 experiments, we compared variations on the game interface, use of the tutoring system, and the form of the feedback. Our findings suggest that for learning basic intercultural communicative skills, a 3-dimensional (3-D) interface with animation and sound produced equivalent learning to a more static 2-D interface. However, learners took significantly longer to analyze and respond to the actions of animated virtual humans, suggesting a deeper engagement. We found large gains in learning across conditions. There was no differential effect with the tutor engaged, but it was found to have a positive impact on learner success in a transfer task. This difference was most pronounced when the feedback was delivered in a more general form versus a concrete style.

Keywords: virtual humans, intelligent tutoring systems, sense of presence, feedback, intercultural communication

Pedagogical agents are animated characters that inhabit virtual learning environments and usually play the role of tutor (Haake & Gulz, 2009; Johnson, Rickel, & Lester, 2000) or peer (Y. Kim & Baylor, 2006). In these roles, the agent typically works alongside the learner to provide guidance (Arroyo, Woolf, Royer, & Tai, 2009), hold conversations (Graesser & McNamara, 2010), and encourage and motivate (Baylor, 2011), among many other forms of possible scaffolding. The role of pedagogical agents in virtual learning environments continues to expand. One use of pedagogical agents is replacing a human role player. Thus, instead of the agent assisting the learner with problems, it is the interaction itself with the agent that is intended to have educational value. Here, the agent is usually a *virtual human* playing a defined social role, with learners also playing a role and using specific communicative skills to achieve goals. For example, to prepare for an international business trip, a learner might meet with a virtual foreign business partner from the country of interest to negotiate a fictional contract agreement.

The technology challenge is to simulate social encounters in realistic ways and in authentic contexts. The pedagogical challenge

is to design scenarios in ways that achieve the learning goals, maintain a high level of real-world fidelity, and stay within an ideal window of challenge (whatever that may be). The basic problems of doing this with virtual humans are eloquently stated by Gratch and Marsella (2005):

These “virtual humans” must (more or less faithfully) exhibit the behaviors and characteristics of their role, they must (more or less directly) facilitate the desired learning, and current technology (more or less successfully) must support these demands. The design of these systems is essentially a compromise, with little theoretical or empirical guidance on the impact of these compromises on pedagogy. (p. 256)

The natural tendency is to build simulations to maximize realism since authentic practice opportunities are essential both for learner motivation and transfer to real-world contexts (Sawyer, 2006). However, some questions have been raised regarding the definition of realism as it applies to human communicative behaviors. Human variability due to personality and cultural differences suggest that virtual humans may have a small amount of flexibility to adapt to learners’ needs while remaining realistic (Wray et al., 2009). Further, the design of virtual human scenarios can have a profound influence on the efficacy of the resulting learning experiences and should be carefully constructed to exercise the targeted communicative skills (Ogan, Alevan, Jones, & Kim, 2011).

In this article, we describe a game-based system for teaching intercultural communication skills and an associated intelligent tutoring system (ITS). We then present two studies investigating issues related to *fidelity* and *feedback*, both of which are important factors in virtual learning environments with virtual role players. The goal is to identify the influences of these factors on learner behaviors and on their acquisition of new communication skills. The article ends with a summary of the results, limitations of our studies, and a discussion of future research topics.

This article was published Online First September 9, 2013.

H. Chad Lane, Matthew Jensen Hays, Mark G. Core, and Daniel Auerbach, Institute for Creative Technologies, University of Southern California.

The project described here was sponsored by the U.S. Army. Statements and opinions expressed do not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred. We thank Julia Kim, Eric Forbell, and the BiLAT team for their contributions. We also thank Bob Wray and Brian Stensrud at SoarTech for their support.

Correspondence concerning this article should be addressed to H. Chad Lane, University of Southern California–Institute for Creative Technologies, 12015 Waterfront Drive, Los Angeles, CA 90094. E-mail: lane@ict.usc.edu

The Acquisition of Intercultural Communication Skills

Social skills (or equivalently, interpersonal skills) form the foundation for both simulation of communicative skills (using a virtual human) and for teaching communicative competence. Although no clear consensus has emerged on a single definition of social skills, most include the notions of choosing *appropriate* and *effective* communicative actions for a given context (Segrin & Givertz, 2003). Because of our specific focus on intercultural communication, we adopted the more precise definition of social skills as “the ability of an interactant to choose among available communicative behaviors in order that [she or] he may successfully accomplish [her or] his own interpersonal goals during an encounter while maintaining the face and line of his fellow interactants” (Wiemann, 1977, p. 198). It is worth noting that what constitutes success in a social interaction is not always obvious, rational, or consistent. Further, how interpersonal and communicative goals are established may or may not be evident (Spitzberg & Cupach, 2002).

Despite the peculiarities of human communication, the concept of social skills can be broken down in many different ways. One of the simplest is to consider two fundamental processes: *message reception* (Wyer & Adaval, 2003) and *message production* (Berger, 2009). Message reception refers to one’s ability to both interpret social signals of others (such as speech and nonverbal behaviors) and infer meaning from the communicative acts conveyed by those social signals. The receiver must both have (a) the motivation to interpret and process the message and (b) the knowledge necessary to comprehend it (Wyer & Adaval, 2003).

Challenges to successful decoding of a message can come from contextual and pragmatic sources in the immediate environment, as well as from internal biases or beliefs. For example, assumptions one makes on the basis of stereotypes can greatly impede message reception. On the message production side, similar challenges arise. How one forms a message (consciously or not) depends again on context, beliefs, biases, and so on. Automated communicative skills are deeply rooted and, thus, difficult to modify in ways that enhance the odds of producing more effective outgoing messages. Nonetheless, the acquisition of novel communicative skills has been shown to follow the same patterns as other cognitive skills (Greene, 2003), and so the same techniques used to promote learning should apply. For example, it is known that repeated practice opportunities with feedback are an essential component in the development of expertise (J. R. Anderson, Corbett, Koedinger, & Pelletier, 1995; Kluger & DeNisi, 2004; Shute, 2008). We have applied these foundational principles in our work by providing a virtual practice environment for intercultural communication skills with automated feedback.

Virtual Humans as Role Players

Live role playing has a long history in education (Kane, 1964) and for teaching social interaction skills (Mendenhall et al., 2006; Segrin & Givertz, 2003). There are problems, however, with the approach. First, role playing in classrooms is not situated in a realistic context, which potentially limits transfer of the learned skills. Second, when peers act as role players, the attitudes, conversational content, and so forth of the role play may not be authentic or realistic. Third, expert human role players are generally believed to be the best option but are not cost effective and can

be prone to inconsistency and fatigue. Although virtual humans have significant limitations, they undoubtedly address some of these complex issues (Cassell, Sullivan, Prevost, & Churchill, 2000; Lim, Dias, Aylett, & Paiva, 2012).

Empirical Support for Learning With Virtual Humans

Can virtual humans be effective role-players? Seminal work presented by Reeves and Nass (1996) in *The Media Equation* showed that people bring many of their usual assumptions about human–human interaction to computer-based interactions. Further, evidence is mounting that this result holds even more strongly when the computer presents a virtual agent (Gratch, Wang, Gerten, Fast, & Duffy, 2007; Pfeifer & Bickmore, 2011; Zambaka, Ulinski, Goolkasian, & Hodges, 2007). In other words, people treat virtual humans as if they are real. Further, characters who provide *personalized* interactions are known to increase feelings of social presence, which in turn enhance learning (Moreno & Mayer, 2004). Learning can also be enhanced when learners choose to adopt social goals (e.g., “come to know your partner”) while interacting with virtual humans (Ogan, Kim, Alevan, & Jones, 2009). Together, these results suggest that virtual humans can induce feelings of social presence in learners, that these feelings are enhanced through personalization and simulation of social and relational behaviors, and, ultimately, that we should expect a concomitant improvement in learning.

Early studies of the efficacy of virtual-human-based systems to teach intercultural skills seem to support this conclusion. Significant gains in overall learning were found for Tactical Iraqi (Surface, Dierdorff, & Watson, 2007) as well as Bilateral Negotiation Trainer (BiLAT; Durlach, Wansbury, & Wilkinson, 2008; J. M. Kim et al., 2009; Lane, Hays, Auerbach, & Core, 2010). Unfortunately, these and similar studies of other virtual learning environments for culture do not compare the systems with traditional (e.g., classroom-based) intercultural training, so it is not yet known if they are more effective than classroom-based learning (Ogan & Lane, 2010).

Virtual humans have been used successfully as role players in many contexts. For example, virtual agents have served as patients in clinical training (Johnsen, Raij, Stevens, Lind, & Lok, 2007), persons of interest in police officer training (Hubal, Frank, & Guinn, 2003), modelers of healthy play for children with autism (Tartaro & Cassell, 2008), victims and perpetrators of bullying in school settings (Aylett, Vala, Sequeira, & Paiva, 2007; Sapouna et al., 2010), and modelers of coping behaviors for mothers of children with serious illness (Marsella, Johnson, & LaBore, 2000). A key question for the intelligent virtual agent community is whether effectiveness will also increase with increased sophistication of the agents.

BiLAT: Teaching Bilateral Negotiation With Cultural Awareness

The context for our work is BiLAT, a game-based simulation for practicing the preparation, execution, and understanding of bilateral meetings in a cultural context (J. M. Kim et al., 2009). As part of an overarching narrative, learners prepare and meet with a series of virtual humans to solve problems in a fictional Iraqi city. Even though BiLAT’s overall scope is much broader, our focus is on

face-to-face meetings between learners and virtual characters and the basic intercultural communicative skills necessary to build trust and reach agreements. BiLAT meetings emphasize both message production and reception skills as discussed earlier.

In BiLAT, learners meet with one or more characters to achieve a set of predefined objectives. For example, the learner may need to convince a high-ranking local official to stop imposing certain taxes on his people or reach an agreement about who will provide security at a local marketplace. In all cases, the learner is required to adhere to Arab business cultural expectations and norms (Nydell, 2006), establish a relationship through building trust, and apply integrative negotiation techniques. Specifically, BiLAT is designed as a practice environment for learning win/win negotiation techniques, which promotes the idea that negotiation counterparts should proactively strive to meet each other's needs as well their own (Fisher, Ury, & Patton, 1991). To achieve this in BiLAT, learners must also apply their understanding of the character's culture to modify their own communicative choices (Landis, Bennett, & Bennett, 2004). BiLAT's focus is on the communicative intent and the structure of meetings and does not seek to teach new languages.

Screenshots of the BiLAT interface are shown in Figure 1. On the left is one of several navigation screens used in the game. On the right is the meeting screen, where learners spend much of their time during play. To communicate with the virtual character (i.e., apply message production skills), the learner selects from a menu of about 70 conversational actions that can vary between scenarios. For example, the learner can engage in small talk (e.g., "talk about soccer"), ask questions (e.g., "ask who is taxing the market" and "ask if he enjoys travel"), and state intentions (e.g., "say you are interested in finding a mutually beneficial agreement"), among other possibilities. Physical actions are also available (e.g., "remove sunglasses" or "give medical supplies"). Corresponding dialogue text is displayed in a dialogue pane while the character responds with synthesized speech and animated gestures.

BiLAT characters possess culturally specific models of how they expect meetings to progress. This progression includes expectations for an opening phase, a social period, a business period, and a closing social period. These phases are derived from live role playing sessions and cognitive task analysis performed with subject-matter experts early in the development of BiLAT (J. M. Kim et al., 2009). An example of a knowledge component taught by BiLAT is to *follow the lead of your host*. If a learner chooses an action that is not appropriate for the current phase of a meeting, the character will respond negatively. Trust, which is directly

affected by the ability of the learner to take appropriate and effective actions, is a major factor in whether BiLAT characters will be agreeable or difficult. When trust is not established, it is often impossible to achieve all necessary agreements because the character will not be as interesting in working together. This means that learners often need multiple follow-up meetings with the same character to achieve objectives and to try different strategies for building trust.

Intelligent Tutoring in BiLAT

The intelligent tutoring system in BiLAT provides feedback to learners as they interact with characters. It is based on knowledge components that were identified during the initial cognitive task analysis and uses them to maintain a learner model and generate the content for feedback messages (Lane et al., 2008). Help can come in the form of *feedback* about a previous action (e.g., explain a reaction from the character by describing an underlying cultural difference) or as a *hint* about what action is appropriate at the given time. Both types of messages appear in the BiLAT dialogue pane (shown in Figure 1). Further, the system implements an adjustable model-scaffold-fade algorithm that reduces coaching support with increased time and successful interactions (Collins, Brown, & Newman, 1989).

Assessment of learner actions is driven by a model of intercultural interactions for Arab business meetings. We defined a typing system for the lowest level elements in the knowledge component hierarchy to facilitate the ITS' understanding of the different categories of message production. These include *required steps*, *usual steps* (but not required), *rules of thumb*, and *avoids* and are identified as tags on steps in recipes for achieving certain communication or negotiation goals. For example, the knowledge components include recipes for greeting, socializing, eliciting the perspective of the counterpart, asking about local events, and more. Which tags belong in which recipes was completed as a joint authoring effort between researchers and subject-matter experts.

These scenario-independent recipes were then mapped into communicative actions available in the game. This allowed the ITS to track learner actions in terms of knowledge components and evaluate actions as positive or negative instances of understanding those components. This authoring task was also performed jointly between researchers and subject-matter experts. It was often necessary to assign two links to some actions that had both positive and negative elements. For example, if a learner promises to give a character what she or he wants, the relationship with that char-



Figure 1. Screenshots of Bilateral Negotiation Trainer, a serious game for intercultural communication.

acter may be enhanced, but the promise could lead to problems down the road (e.g., the character's neighbors may grow jealous and demand the same favors). These trade-offs were highlighted by the ITS when they occurred—there were usually reasons to take the action (or respond) and reasons not to do it, and the best choice depended on the payoff and specific problem being solved.

Measuring Learning From BiLAT and the ITS

In the experiments described, we used two measures to evaluate learning produced by BiLAT and the ITS. The first measure was a *situational judgment test* (SJT). In general, SJTs present several domain-relevant scenarios, each of which is accompanied by several actions that the learner might perform in response to the scenario (Legree & Psotka, 2006). The participants provided Likert-scale ratings for each action (0 = *very poor action*, 5 = *mixed/OK action*, 10 = *very good action*). There were eight total scenarios and 28 total actions in the SJT (these items were provided by an external team at the U.S. Army Research Institute). The following is an example situation and to-be-rated actions:

Major Cross and Hamad are wrapping up their meeting, right on schedule. There are only a few minutes left in the allotted time for the meeting. Before the meeting, Hamad explained that he would need to leave at a particular time so that he is able to get to the mosque in time for afternoon prayer. Rate the following ways in which Major Cross could end the meeting.

(0–10) ___ Revisit any results of the meeting that were unsatisfactory and try to work them out.

(0–10) ___ Make sure Hamad clearly understands all agreements. If the meeting runs a little longer than scheduled, it is okay.

(0–10) ___ Spend some social time together and remind Hamad that his friendship is valuable.

To score the participant responses, we used ratings provided by three subject-matter experts. Understanding of the domain knowledge is defined as the degree that a participant's responses correlate with the experts' responses (Legree & Psotka, 2006). The test was administered in a pretest–posttest design, and so learning was defined as the increase in the correlation from pretest to posttest. Because the situational judgment test focused on the participants' ability to recognize and understand concepts about intercultural interaction, it measured learning at the lower levels of Bloom's taxonomy of cognitive skills (L. W. Anderson & Krathwohl, 2001).

The second measure was an in-game posttest that focused on a new issue (supply thefts from an Iraqi hospital rather than the market). During the participants' meetings with a hospital administrator, no feedback was provided. For each action that a participant selected during these meetings, we examined the probability that it was inappropriate. Participants who made fewer errors were said to have learned more about intercultural interaction than were participants who made more errors. We also examined the probability that the participant was able to successfully negotiate with the hospital administrator. Although it was a binary measure, success indicated that the participants were able to build up trust and consider their partner's needs effectively. Because the in-game posttest measured the participants' ability to apply what they learned about intercultural interaction, it measured learning at the

middle levels of Bloom's taxonomy of cognitive skills (L. W. Anderson & Krathwohl, 2001).

Experiment 1: Fidelity and Presence

Although not the only method, one approach to measuring engagement is by investigating to what a degree a system can establish a *sense of presence*. One way to induce a sense of presence is to provide greater visual and auditory fidelity (e.g., more realistic graphics and sound; Lombard & Ditton, 1997). Intuitively, it seems that greater sensory fidelity should also promote better training; this is a common point of emphasis in training system design requirements. However, recent studies on the effect of presence on training suggest that engagement and effective outcomes are enhanced by greater sensory fidelity, but learning does not necessarily improve (Rowe, Shores, Mott, & Lester, 2011). An exception is the case in which a specific task domain requires high-sensory-fidelity simulation (e.g., a flight simulator), but most systems with greater sensory fidelity are not necessarily better trainers as a result.

Experiment 1 was designed to determine whether a *social* simulator must have high visual and auditory fidelity in order to effectively engage and instruct. We therefore created two versions of the system. Both versions had the high *social* fidelity of the standard BiLAT experience: rich characters, extensive dialogue, intricate character backgrounds, and so forth. But only one version had the rich visual and auditory experience; the other used a static, primarily text-based interface. On one hand, because BiLAT is essentially a social-skills trainer, the difference in visual and auditory fidelity may not have affected either presence or learning. On the other, given the tendency of people to treat virtual human interactions as being real (Gratch et al., 2007), we anticipated some advantages for the high fidelity version of BiLAT, including a deeper sense of presence and realism.

Method

Participants. The participants were 46 U.S. citizens (recruited from college campuses) who received \$60 as compensation for approximately 3 hr of participation.

Measures. We used the SJT in a pretest–posttest design, as described previously. We also used the in-game posttest described and analyzed, for each participant, the number of actions they took and the amount of time they deliberated between actions. Participants who took more actions and deliberated for less time were thought to be less engaged or to be taking the experience less seriously.

We added a new measure to capture how engaged the participants were while playing BiLAT: the Temple Presence Inventory (TPI). The TPI is a validated battery of self-report Likert-scale ratings that attempt to measure how engaged or immersed one is in a multimedia experience (Lombard & Ditton, 1997). We used two subscales from the TPI: the *Social* subscale and the *Spatial* subscale. An example of a Social subscale item is "How often (1 = *never*, 7 = *always*) did it feel as if someone you saw/heard in the environment was talking directly to you?" An example of a Spatial subscale item is "How much (1 = *not at all*, 7 = *very much*) did it seem as if you could reach out and touch the objects or people you saw/heard?" (Items on the Spatial subscale that addressed sound or animation were removed.)

Design and procedure. After providing consent, the participants completed the pretest SJT (online, administered by a survey-hosting website). Within a few days, the participants arrived at our Institute and were randomly assigned to one of the two between-subjects conditions. In the three-dimensional (3-D) condition, the participants played BiLAT with the rich, immersive interface previously described. In the 2-D condition, the participants played BiLAT with a nonimmersive, static, text-focused interface (shown in Figure 2). The 2-D interface had neither animation nor sound but was otherwise equivalent to the 3-D interface. That is, the characters and coach functioned identically in both conditions. The participants then completed the in-game posttest (using the same interface as they used with the market scenario). They then completed the two subscales of the TPI described previously. Finally, they completed the posttest SJT.

Results

Presence and immersion. The primary results from Experiment 1 are split into Table 1 and Table 2. Table 1 presents the participants' self-reported presence, the number of meetings they conducted with each character, and the number of actions they took in each meeting.

As can be seen, there was a main effect of interface on self-reported presence. The 3-D interface yielded higher spatial presence ratings than did the 2-D interface: $t(44) = 3.091, p = .003$, partial $\eta^2 = .178$. The 3-D interface also yielded higher social presence ratings than did the 2-D interface: $t(44) = 2.542, p = .015$, partial $\eta^2 = .128$.

There was also a main effect of interface on how the participants interacted with the virtual characters. (A software error corrupted the logs for two participants. Their data did not contribute to this analysis.) The participants conducted more meetings in the 2-D interface than in the 3-D interface: $t(42) = 3.143, p = .003$, partial $\eta^2 = .190$. During each meeting, the participants performed more actions in the 2-D interface than in the 3-D interface: $t(42) =$

$2.546, p = .015$, partial $\eta^2 = .134$. Summed across meetings, participants performed nearly 50% more actions in the 2-D interface than in the 3-D interface in approximately the same amount of time. A similar pattern of results appeared in the in-game posttest. The 3-D interface, it appears, caused people to think more about their actions than did the 2-D interface.

Learning. Table 2 presents the participants' SJT scores and their performance on the in-game posttest.

Declarative knowledge. A repeated-measures mixed analysis of variance (ANOVA) revealed that there was not a main effect of interface on the participants' pretest–posttest gain: $F(1, 44) < 1, ns$. A main effect was not obscured by pretest differences between the two conditions; participants assigned to the 2-D interface did not score reliably higher than those assigned to the 3-D interface: $t(44) = 1.330, p = .191$. Thus, although the 3-D interface created more presence, it did not produce learning gains at the lower levels of Bloom's taxonomy (L. W. Anderson & Krathwohl, 2001).

Application and transfer. There was also not a main effect of interface on the probability of successful negotiation during the in-game posttest: $F(1, 44) = 1.208, p = .278$. Finally, there was not a main effect of interface on the probability of making an error during the in-game posttest: $t(40) = 1.536, p = .132$. Along with the SJT data, it seems clear that greater visual fidelity—and the spatial and social immersion it generates—does not appear to have a substantial effect on learning cross-cultural interactions as addressed in BiLAT.

Overall gains. We conducted additional analyses of the SJT data to examine the overall gains produced by interacting with BiLAT and the ITS. A repeated-measures ANOVA revealed that there was a main effect of practice on the improvement from pretest to posttest. Correlation with subject-matter experts increased from pretest ($M = 0.56, SE = 0.03$) to posttest ($M = 0.72, SE = 0.08$): $F(1, 44) = 40.039, p < .001$ partial $\eta^2 = .476$. Overall, it appears that BiLAT and the ITS can effectively increase knowledge about how to interact in a intercultural context.

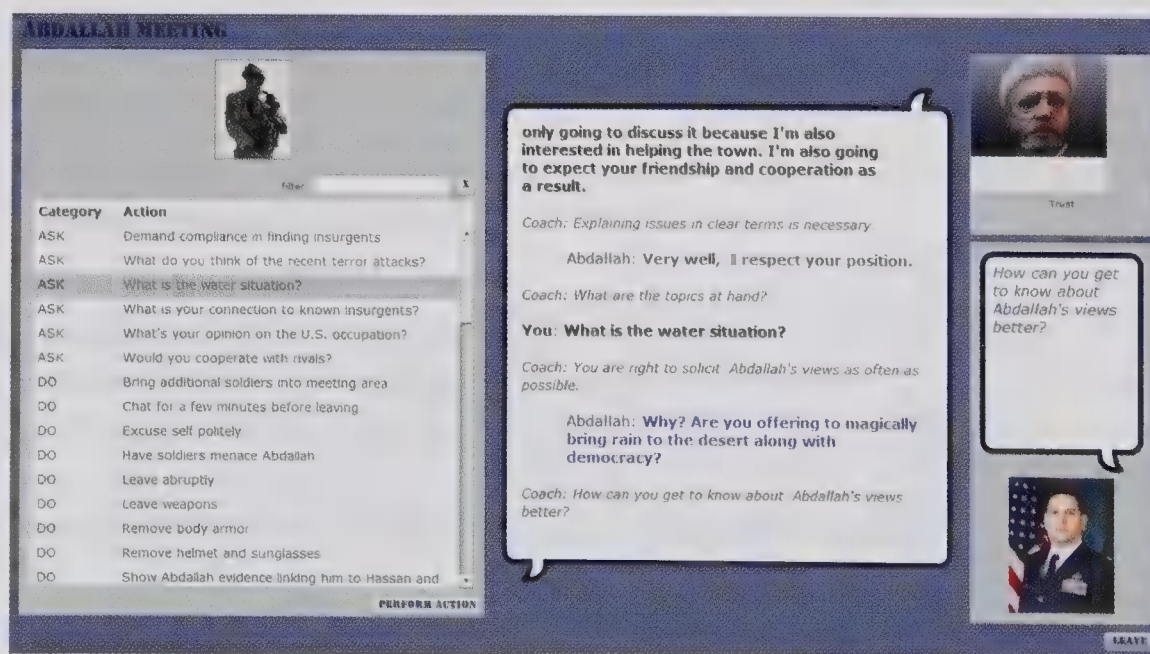


Figure 2. A screenshot of the two-dimensional interface for Bilateral Negotiation Trainer.

Discussion

The 3-D version of BiLAT (with animated virtual humans and synthesized speech) produced a greater sense of presence than did the 2-D interface. However, according to the SJT, there were no differences in declarative knowledge gains between the two conditions. Thus, the 3-D interface did not appear to improve BiLAT's teaching efficiency.

However, there were several differences in the how users in the two conditions interacted with the characters. Learners in the 3-D environment deliberated longer and, correspondingly, needed fewer actions in order to succeed. Research on rapport with virtual humans has shown that people react to virtual humans as if they are real (Gratch et al., 2007). One possible explanation for the interface-driven behavioral differences is that learners were more concerned about the impacts of their choices and thus thought them through more carefully. They may have been using that time to generate better mental simulations of the conversation. They may also have been establishing better expectations or generating better hypotheses about the mental state of their meeting partner. Future studies would be necessary to determine why users deliberate longer with embodied characters and how they are using that time.

Experiment 2: Formative Feedback

The results of Experiment 1 suggested that the content—not the appearance—of the system appeared to be responsible for learning. We therefore designed Experiment 2 to focus on the effectiveness of that content by examining the hints and feedback provided by the ITS. Some participants received formative feedback (Shute, 2008), which emphasizes productive revisions of knowledge. This feedback was very conceptual in nature (e.g., “Be sure to avoid appearing overly defensive or protective”). Other participants received very helpful, but very specific, assistance (e.g., “You are still in full combat gear, including your helmet and sunglasses”).

Prior studies have found that learners who struggle during training eventually prosper as a result (Bjork, 1994; VanLehn, 1988). We believed that the specific feedback would be more helpful during practice and be appealing to learners (since it told them exactly what to do), but that the conceptual feedback would require the participants to deliberate more and think more deeply

Table 1
Temple Presence Inventory and Meeting–Interaction Data From Experiment 1 by Condition

Interface	TPI				No. of interactions			
	Social		Spatial		Meetings		Actions	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
2-D	2.77	0.19	2.30	0.20	13.67	0.91	17.70	0.90
3-D	3.49	0.18	3.21	0.21	10.30	0.60	15.09	0.54

Note. Participants' self-reported presence, the number of meetings they conducted with each character, and the number of actions they took in each meeting. TPI = Temple Presence Inventory; Social = Social subscale of the TPI; Spatial = Spatial subscale of the TPI; SE = standard error; 2-D = two-dimensional interface; 3-D = three-dimensional interface.

Table 2

Situational Judgment Test and In-Game Posttest Data From Experiment 1 by Condition

Interface	SJT				In-game posttest			
	Pretest		Posttest		<i>p</i> (success)		<i>p</i> (error)	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
2-D	.59	.04	.72	.03	.37	.11	.32	.02
3-D	.52	.04	.72	.03	.35	.10	.27	.02

Note. Participants' Situational Judgment Test (SJT) scores and their performance on the in-game posttest. 2-D = two-dimensional interface; 3-D = three-dimensional interface; SE = standard error.

about the principles of cross-cultural interaction. We therefore predicted a greater increase in declarative knowledge (greater SJT improvement) and better transfer to new contexts (greater in-game posttest performance) for participants in the formative feedback condition.

Method

Participants. The participants were 47 U.S. citizens (recruited from college campuses) who received \$60 as compensation for approximately 3 hr of participation.

Design and procedure. After providing consent, the participants completed the pretest SJT online (cf. Experiment 1). Within a few days, the participants arrived at our institute and were randomly assigned to one of the two between-subjects conditions. Some of the participants used BiLAT with a coach that provided hints and feedback that were exclusively *specific* to in-game actions. The coach for the other participants provided hints and feedback that were exclusively *conceptual*. The two versions of the coach¹ otherwise behaved identically in the two conditions (e.g., they chose when to provide feedback or hints based on the same policies).

The participants then completed the in-game posttest. They were then compensated and dismissed. A week later, the participants were e-mailed a link to the posttest SJT; 46 of the 47 participants completed it after an average of about 2 days.

Results

The primary results from Experiment 2 are presented in Table 3.

Declarative knowledge. There was not a main effect of feedback type on the participants' pretest–posttest gain: $F < 1$, *ns*. Their acquisition of declarative knowledge appears to not have been influenced by specific versus conceptual feedback.

Application and transfer. On the in-game posttest, there was not a main effect of coach type on the probability of a successful meeting outcome: $t < 1$, *ns*. However, there was a main effect of feedback type on the probability of making an error: $t(40) = 2.049$, $p = .05$, partial $\eta^2 = .095$. Even with equivalent declarative knowledge, the participants who encountered the conceptual coach were better able to interact with the new character in order to solve

¹ Typically, the coach follows a simple policy to decide whether to provide specific or conceptual feedback during practice (which is to try general first and then shift to concrete if the learner struggles).

Table 3
Effects of Feedback Type on Situational Judgment Test Scores and In-Game Posttest Performance

Coach type	SJT				In-game posttest			
	Pretest		Posttest		<i>p</i> (success)		<i>p</i> (error)	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Specific feedback	.58	.04	.75	.02	.77	.09	.20	.02
Conceptual feedback	.56	.04	.74	.02	.76	.10	.15	.02

Note. SJT = Situational Judgment Test; SE = standard error.

the problem. This pattern of results is consistent with the notion that formative feedback can be superior to simple performance-oriented feedback (Shute, 2008). Further, the disparity between the in-game posttest results and the SJT is consistent with our belief that these two measures operated at different levels of Bloom’s taxonomy of learning (L. W. Anderson & Krathwohl, 2001) taxonomy (Hays, Ogan, & Lane, 2010).

Overall gains. On the SJT, there was a main effect of practice with BiLAT. A repeated-measures ANOVA revealed that the participants’ SJT scores increased from pretest to posttest: $F(1, 44) = 61.169, p < .001$, partial $\eta^2 = .582$. As in Experiment 1, it appears that the participants learned from their in-game experience.

Discussion

In Experiment 2, we tested the hypothesis that general feedback would be better for learning. The results suggested that conceptual feedback transfers more readily than does concrete feedback. Although we cannot conclude that concrete feedback never has a place (extreme versions of the ITS were used in the experiment), this study does suggest that for intercultural communication skills, a reasonable default setting is to use conceptual feedback first and then shift to concrete if future performance gains are not observed.

General Discussion

We sought to build a virtual environment for teaching intercultural communication skills with virtual humans in a specific context (i.e., Arab business practice). Our approach included the use of modern game technologies (i.e., a 3-D game engine) and an intelligent tutor to scaffold learners as they interacted with those characters. We conducted two experiments in which we found large overall gains in declarative knowledge as a result of interacting with this system. We found that the visual fidelity of the interface had significant impacts on learner behaviors, perceptions, and in-game success. We also found that conceptual feedback enhanced learners’ ability to apply the targeted knowledge. At least in the context of intercultural communication, the nature of feedback had a much greater influence on learning than did visual fidelity.

Fidelity

For social simulations and virtual humans, the choice of where to invest development time is challenging: It is difficult to ignore any single dimension but also difficult to develop elaborate models

for all relevant aspects of the communicative skill. Our studies suggest that for the message reception/production model of communication with a menu-based simulation of social interactions, learning of declarative knowledge is not affected by the richness of the sound and animations. However, given the complexities of social interactions, there are analogs to other domains that require higher fidelity simulations, like flight training. For example, virtual human agents used for teaching recognition of nonverbal behaviors (e.g., in deception detection training) would require a higher level of visual fidelity to properly capture and teach the subtle elements that are part of the knowledge being covered.

In BiLAT, learners are practicing the decision making involved in intercultural communication and learning what differences require attention. Variations in speech and nonverbal behaviors are not as critical, given these goals, and so the fidelity important to BiLAT learning has most to do with the content of the characters utterances, which is driven by the underlying models. The result that a richer interface engendered longer deliberations suggests that future studies are needed in order to understand the nature of how this time is being used: if virtual humans and high-fidelity graphics can be linked to greater attention to consequences of actions or more self-explanations, then future studies should seek to determine if these do in fact contribute to learning.

Feedback

Presence is often defined as forgetting that one is having a mediated experience. Thus, it is important to understand (a) if the use of unsolicited feedback interrupts this experience and (b) if that positively or negatively impacts learning. We found no evidence that the use of feedback (from a disembodied ITS) impacted the learner’s sense of presence in either environment. How to deliver feedback optimally is an ongoing question for learning science researchers. Our study found benefits for using more general feedback that, we posit, required the learner to interpret the help and apply it to his or her own situation (i.e., it is *formative*). As our study only tested the extremes, an ITS that properly balances concrete feedback with conceptual feedback is more likely to be effective for the most learners. Future studies should focus on various algorithms for comparing different timing and content settings.

Limitations and Future Work

There were several technical and methodological limitations of the present studies.

Modality. Since communicating with BiLAT characters is accomplished through menu-based selection of actions rather than free speech input, learners are limited in what they can say and are most likely influenced by the choices that are available. This is vastly different than being required to generate utterances as they would in normal conversation. On the other hand, this design choice reflects current limitations of speech input and natural language understanding and does provide some structure for novice learners (J. M. Kim et al., 2009). Thus, because our measures focus on culture at the same level of abstraction as the game, it is unclear whether BiLAT practice with coaching would transfer to more realistic contexts. Because this is a critically important question, it suggests further study using a more elaborate (and expen-

sive) a posttest measure using human role players or perhaps virtual humans that are capable of understanding free speech input.

Feedback. As discussed, feedback in BiLAT is delivered as text. An important lesson learned about the delivery of feedback resulted from early testing when we discovered that many learners were not reading the coaching messages. The reason, we found out, was that the BiLAT display included a readout of how much “trust” the character felt toward the learner. Thus, people played by selecting an action, *listening* to the character’s response, and watching for changes in the trust meter. Because the messages were rendered only as text in the dialogue pane (on the other side of the screen from the trust meter), they often went unseen. We resolved this issue by hiding the trust meter in all of our studies and drawing attention to the coach and how it worked when introducing the participants to the system. It may be that having a trust meter or other visible “score” may have benefits for learning (e.g., rapid self-assessments), and thus our studies are limited in that they do not address the interaction of learning and game-like elements, like score, narrative, or challenge.

Further, our studies focus only on limited forms of feedback delivery. In Experiment 2, we considered two extreme versions of feedback—either completely general or completely concrete. It is likely that some balance between these is needed, and thus, continued study of how the generality of specificity of feedback should vary with context is needed. Also, the settings on our model–scaffold–fade algorithm were based on trial-and-error and observation of learners interacting with the system. Our goal was to strike a balance and provide the best level of support given the successes (or failures) of the learner, yet we had no theoretical support for the settings we used on this algorithm. Although there is substantial literature on the form of feedback (Shute, 2008), in general, we have found little guidance—empirical or theoretical—regarding the timing and optimal rates of fading of tutor support. This suggests future studies varying our algorithm and investigating the impact of these on performance and learning.

Measures. Unfortunately, both of the measures we used to gauge changes in learners have drawbacks. Although the in-game posttest does detect changes in learners’ understanding of some culturally based rules of interaction, it is conducted within the environment used to teach those rules. As a result, it may only reflect shallow learning (i.e., learning to play the game rather than learning about the culture) or basic evidence of the existence of smooth learning curves. Because the ultimate test of learning is in face-to-face interactions with people from the target culture, in-game performance measures are inherently suggestive, at best.

Also, as discussed, although the SJT was designed by an external team, it may not be sufficiently precise to detect learning that occurs during BiLAT meetings. Further, it includes content from components of BiLAT that are not part of the tutoring system, such as preparation (i.e., research on counterparts) and broader scenario issues, like following up on commitments and social network changes based on the overarching narrative in the scenarios. It should also be noted that both these measures focus exclusively on the *message production* side of social interaction. Thus, even though the ITS does address *message reception* skills, our studies had no chance to detect any changes in a learner’s ability to process and understand the utterances from the virtual human characters.

Another potential limitation of the SJT is that it uses identical prompts on the pretest and posttest. One could therefore argue that the overall gains from pretest to posttest we have reported merely reflect learning from the test. We took care to reduce this possibility by modifying some of the prompts to avoid divulging additional information (Asher, 2007). Also, because the SJT responses are numeric rather than potentially informative solutions (as in multiple-choice tests), it is unlikely that the participants used the SJT to guide their BiLAT experience so that their posttest score would be improved. Nevertheless, multiple counterbalanced versions of the SJT would be a more empirically sound measure.

Many other measures were possible, such as perspective-taking instruments (Paige, 2006) and measures to gauge perceived relationships with the agent (Ogan et al., 2011). Such measures are extremely important in intercultural development because much of it involves adjustment of one’s own perspective on self, others, and more (Bennett, 1993), and so in future studies, investigators should more carefully address the role of feedback and fidelity on these factors while respecting the practical limits on testing time used during controlled studies.

Conclusion

This article began with the question of how virtual human role players might be used to enhance the learning of communication skills and highlighted the dearth of guidelines, principles, and empirical evidence for their design. Broadly, the results of our studies support the limited, but growing, body of literature (Durlach et al., 2008; Surface et al., 2007) that virtual humans can be used effectively to improve intercultural communication knowledge and skills. Generally, learners in both of our experiments showed gains in declarative knowledge from pretest to posttest. The key takeaway messages from these studies are that (a) the fidelity of such systems should be a function of the domain knowledge being taught and (b) feedback can be given in such a way that it enhances future performance and does not distract from the immersive nature of the system. Although our studies were not specifically “design” studies, further investigation of precise manipulations of virtual human content, behavior, and interaction modalities is definitely necessary. As with many advanced technologies (games, mobile devices, and so on), the number of available systems from the commercial and research sectors is rapidly growing, and so there is an urgent need for empirically derived guidelines and principles for using and scaffolding learning with virtual humans.

Many open questions remain about the use of virtual humans in social skills training and education. We believe future work is needed to develop and test new measures of learning and perspective change and to understand the role of feedback in these environments. As virtual humans continue to approach live human role players in realism, continued experimental research that focuses on the nature of these interactions, the sophistication of their implementations, and the role of supporting technologies such as intelligent tutoring is certainly merited.

References

- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4, 167–207. doi:10.1207/s15327809jls0402_2

- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational outcomes*. New York, NY: Longman.
- Arroyo, I., Woolf, B. P., Royer, J. M., & Tai, M. (2009). Affective gendered learning companions. In V. Dimitrova, R. Mizoguchi, B. Du Boulay, & A. C. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 41–48). Amsterdam, the Netherlands: IOS Press.
- Asher, H. (2007). *Polling and the public: What every citizen should know* (7th ed.). Washington, DC: CQ Press.
- Aylett, R., Vala, M., Sequeira, P., & Paiva, A. (2007). FearNot! An emergent narrative approach to virtual dramas for antibullying education. In M. Cavazza & S. Donikian (Eds.), *Virtual storytelling: Using virtual reality technologies for storytelling* (Lecture Notes in Computer Science Vol. 4871, pp. 202–205). Berlin, Germany: Springer. doi:10.1007/978-3-540-77039-8_19
- Baylor, A. (2011). The design of motivational agents and avatars. *Educational Technology Research and Development*, 59, 291–300. doi:10.1007/s11423-011-9196-3
- Bennett, M. J. (1993). Toward ethnorelativism: A developmental model of intercultural sensitivity. In R. M. Paige (Ed.), *Education for the intercultural experience* (2nd ed., pp. 21–71). Yarmouth, ME: Intercultural Press.
- Berger, C. R. (2009). Message production processes. In C. R. Berger, M. E. Roloff, & D. R. Roskos-Ewoldsen (Eds.), *The handbook of communication science* (2nd ed., pp. 111–128). Thousand Oaks, CA: Sage.
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.
- Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (Eds.). (2000). *Embodied conversational agents*. Cambridge, MA: MIT Press.
- Collins, A., Brown, J. S., & Newman, D. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction* (pp. 453–494). Mahwah, NJ: Erlbaum.
- Durlach, P. J., Wansbury, T. G., & Wilkinson, J. (2008). *Cultural awareness and negotiation skill training: Evaluation of a prototype semi-immersive system*. Retrieved from the Defense Technical Information Center website at <http://www.dtic.mil/dtic/>
- Fisher, R., Ury, W., & Patton, B. (1991). *Getting to yes: Negotiating agreement without giving in*. New York, NY: Penguin Books.
- Graesser, A., & McNamara, D. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist*, 45, 234–244. doi:10.1080/00461520.2010.515933
- Gratch, J., & Marsella, S. (2005). Lessons from emotion psychology for the design of lifelike characters. *Applied Artificial Intelligence*, 19, 215–233. doi:10.1080/08839510590910156
- Gratch, J., Wang, N., Gerten, J., Fast, E., & Duffy, R. (2007). Creating rapport with virtual agents. In C. Pelachaud, J. Martin, E. Andre, G. Chollet, K. Karpouzis, & D. Pele (Eds.), *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (pp. 125–138). Berlin, Germany: Springer-Verlag.
- Greene, J. O. (2003). Models of adult communication skill acquisition: practice and the course of performance improvement. In J. O. Greene & B. R. Burleson (Eds.), *Handbook of communication and social interaction skills* (pp. 51–91). New York, NY: Routledge.
- Haake, M., & Gulz, A. (2009). A look at the roles of look & roles in embodied pedagogical agents: A user preference perspective. *International Journal of Artificial Intelligence in Education*, 19, 39–71.
- Hays, M. J., Ogan, A. E., & Lane, H. C. (2010). The evolution of assessment: Learning about culture from a serious game. In C. Lynch, K. Ashley, T. Mitrovic, V. Dimitrova, N. Pinkwart, & V. Aleven (Eds.), *Proceedings of the 4th International Workshop on Intelligent Tutoring Technologies for Ill-Defined Problems and Ill-Defined Domains Held at the 10th International Conference on Intelligent Tutoring Systems (ITS 2010) in Pittsburgh, Pennsylvania* (pp. 37–44). Clausthal-Zellerfeld, Germany: Clausthal University of Technology.
- Hubal, R. C., Frank, G. A., & Guinn, C. I. (2003). Lessons learned in modeling schizophrenic and depressed responsive virtual humans for training. In L. Johnson & E. Andre (Eds.), *Proceedings of the 8th International Conference on Intelligent User Interfaces* (pp. 85–92). Miami, FL: ACM Press.
- Johnsen, K., Raij, A., Stevens, A., Lind, D. S., & Lok, B. (2007). The validity of a virtual human experience for interpersonal skills education. In D. Gilmore (Ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1049–1058). San Jose, CA: ACM Press.
- Johnson, W. L., Rickel, J., & Lester, J. C. (2000). Animated pedagogical agents: face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11, 47–48.
- Kane, P. E. (1964). Role playing for educational use. *Speech Teacher*, 13, 320–323. doi:10.1080/03634526409377396
- Kim, J. M., Hill, R. W., Durlach, P. J., Lane, H. C., Forbell, E., Core, M., . . . Hart, J. (2009). BiLAT: A game-based environment for practicing negotiation in a cultural context. *International Journal of Artificial Intelligence in Education*, 19, 289–308.
- Kim, Y., & Baylor, A. (2006). A social-cognitive framework for pedagogical agents as learning companions. *Educational Technology Research and Development*, 54, 569–596. doi:10.1007/s11423-006-0637-3
- Kluger, A. N., & DeNisi, A. (2004). Feedback interventions: Toward the understanding of a double-edged sword. In T. F. Oltmanns & R. E. Emery (Eds.), *Current directions in abnormal psychology* (pp. 76–82). Upper Saddle River, NJ: Pearson Education.
- Landis, D., Bennett, J. M., & Bennett, M. J. (2004). *Handbook of intercultural training*. Thousand Oaks, CA: Sage.
- Lane, H. C., Hays, M. J., Auerbach, D., & Core, M. (2010). Investigating the relationship between presence and learning in a serious game. In J. Kay & V. Aleven (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 274–284). Berlin, Germany: Springer-Verlag.
- Lane, H. C., Hays, M. J., Auerbach, D., Core, M., Gomboc, D., Forbell, E., & Rosenberg, M. (2008). Coaching intercultural communication in a serious game. In S. Chen, A. Mitrovic, & R. Mizoguchi (Eds.), *18th International Conference on Computers in Education* (pp. 35–42). Taipei, Taiwan: Asia-Pacific Society for Computers in Education.
- Legree, P. J., & Psotka, J. (2006, November). *Refining situational judgment test methods*. Paper presented at the Proceedings of the 25th Army Science Conference, Orlando, FL.
- Lim, M. Y., Dias, J., Aylett, R., & Paiva, A. (2012). Creating adaptive affective autonomous NPCs. *Autonomous Agents and Multi-Agent Systems*, 24, 287–311. doi:10.1007/s10458-010-9161-2
- Lombard, M., & Ditton, T. (1997). At the heart of it all: the concept of presence. *Journal of Computer-Mediated Communication*, 3(2).
- Marsella, S. C., Johnson, W. L., & LaBore, C. (2000). Interactive pedagogical drama. In C. Sierra, M. Gini, & J. S. Rosenschein (Eds.), *Proceedings of the 4th International Conference on Autonomous Agents* (pp. 301–308). Barcelona, Spain: ACM Press.
- Mendenhall, M. E., Stahl, G. K., Ehnert, I., Oddou, G., Osland, J. S., & Kuhlmann, T. M. (2006). Evaluation studies of cross-cultural training programs: A review of the literature from 1998 to 2000. In D. Landis, J. M. Bennett, & M. J. Bennett (Eds.), *Handbook of intercultural training* (pp. 129–144). Thousand Oaks, CA: Sage.
- Moreno, R., & Mayer, R. E. (2004). Personalized messages that promote science learning in virtual environments. *Journal of Educational Psychology*, 96, 165–173. doi:10.1037/0022-0663.96.1.165

- Nydell, M. K. (2006). *Understanding Arabs: A guide for modern times*. Boston, MA: Intercultural Press.
- Ogan, A., Aleven, V., Jones, C., & Kim, J. (2011). Persistent effects of social instructional dialog in a virtual learning environment. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education, Auckland, New Zealand* (Proceedings Series: Lecture Notes in Computer Science Vol. 6738, pp. 238–246). Berlin, Germany: Springer-Verlag.
- Ogan, A., Kim, J., Aleven, V., & Jones, C. (2009). Explicit social goals and learning in a game for cross-cultural negotiation. In S. Craig & D. Dicheva (Eds.), *Proceedings of the Workshop on Educational Games at the 14th International Conference on Artificial Intelligence in Education, Brighton, UK* (pp. 51–58). Berlin, Germany: Springer-Verlag.
- Ogan, A., & Lane, H. C. (2010). Virtual environments for culture and intercultural competence. In E. G. Blanchard & D. Allard (Eds.), *Handbook of research on culturally aware information technology: Perspectives and models* (pp. 501–519). Hershey, PA: IGI Global. doi:10.4018/978-1-61520-883-8.ch023
- Paige, R. M. (2006). Instrumentation in intercultural training. In D. Landis, J. M. Bennett, & M. J. Bennett (Eds.), *Handbook of intercultural training* (pp. 85–128). Thousand Oaks, CA: Sage.
- Pfeifer, L. M., & Bickmore, T. (2011). Is the media equation a flash in the pan? The durability and longevity of social responses to computers. In D. Tan, G. Fitzpatrick, C. Gutwin, B. Begole, & W. A. Kellogg (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 777–780). Vancouver, British Columbia, Canada: ACM.
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. New York, NY: Cambridge University Press.
- Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, 21, 115–133.
- Sapouna, M., Wolke, D., Vannini, N., Watson, S., Woods, S., Schneider, W., . . . Aylett, R. (2010). Virtual learning intervention to reduce bullying victimization in primary school: A controlled trial. *Journal of Child Psychology and Psychiatry*, 51, 104–112. doi:10.1111/j.1469-7610.2009.02137.x
- Sawyer, R. K. (2006). The new science of learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 1–16). Cambridge, MA: Cambridge University Press.
- Segrin, C., & Givertz, M. (2003). Methods of social skills training and development. In J. O. Greene & B. R. Burleson (Eds.), *Handbook of communication and social interaction skills* (pp. 135–176). New York, NY: Routledge.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78, 153–189. doi:10.3102/0034654307313795
- Spitzberg, B. H., & Cupach, W. R. (2002). Interpersonal skills. In M. L. Knapp & J. A. Daly (Eds.), *Handbook of interpersonal communication* (pp. 564–611). Thousand Oaks, CA: Sage.
- Surface, E. A., Dierdorff, E. C., & Watson, A. M. (2007). *Special operations language training software measurement of effectiveness study: Tactical Iraqi study final report* (Technical Report No. 2007010602). Raleigh, NC: SWA Consulting.
- Tartaro, A., & Cassell, J. (2008). Playing with virtual peers: Bootstrapping contingent discourse in children with autism. In P. A. Kirschner, J. van Merriënboer, & T. de Jong (Eds.), *Proceedings of the Eighth International Conference for the Learning Sciences* (Vol. 2, pp. 382–389). Utrecht, the Netherlands: International Society of the Learning Sciences.
- VanLehn, K. (1988). Toward a theory of impasse-driven learning. In H. Mandl & A. Lesgold (Eds.), *Learning issues for intelligent tutoring systems* (pp. 19–41). New York, NY: Springer. doi:10.1007/978-1-4684-6350-7_2
- Wiemann, J. M. (1977). Explication and test of a model of communicative competence. *Human Communication Research*, 3, 195–213. doi:10.1111/j.1468-2958.1977.tb00518.x
- Wray, R., Lane, H. C., Stensrud, B., Core, M., Hamel, L., & Forbell, E. (2009). Pedagogical experience manipulation for cultural learning. In S. Craig & D. Dicheva (Eds.), *Proceedings of the Second Workshop on Culturally Aware Tutoring Systems at the 14th International Conference on Artificial Intelligence in Education, Brighton, UK* (pp. 35–44). Berlin, Germany: Springer-Verlag.
- Wyer, R. S., & Adaval, R. (2003). Message reception skills in social communication. In J. O. Greene & B. R. Burleson (Eds.), *Handbook of communication and social interaction skills* (pp. 291–355). Mahwah, NJ: Erlbaum.
- Zanbaka, C., Ulinski, A., Goolkasian, P., & Hodges, L. F. (2007). Social responses to virtual humans: Implications for future interface design. In D. Gilmore (Ed.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1561–1570). San Jose, CA: ACM Press.

Received December 16, 2011

Revision received December 3, 2012

Accepted December 10, 2012 ■

Motivation and Performance in a Game-Based Intelligent Tutoring System

G. Tanner Jackson and Danielle S. McNamara
Arizona State University

One strength of educational games stems from their potential to increase students' motivation and engagement during educational tasks. However, game features may also detract from principle learning goals and interfere with students' ability to master the target material. To assess the potential impact of game-based learning environments, in this study we examined motivation and learning for 84 high-school students across eight 1-hr sessions comparing 2 versions of a reading strategy tutoring system, an intelligent tutoring system (iSTART) and its game-based version (iSTART-ME). The results demonstrate equivalent target task performance (i.e., learning) across environments at pretest, posttest, and retention, but significantly higher levels of enjoyment and motivation for the game-based system. Analyses of performance across sessions reveal an initial decrease in performance followed by improvement within the game-based training condition. These results suggest possible constraints and benefits of game-based training, including time-scale effects. The findings from this study offer a potential explanation for some of the mixed findings within the literature and support the integration of game-based features within intelligent tutoring environments that require long-term interactions for students to develop skill mastery.

Keywords: educational games, intelligent tutoring, motivation and performance

Intelligent tutoring systems (ITSs) are automated tutoring environments that adapt to users based on various well-established cognitive principles and algorithms (Anderson, 1982). This approach has been highly successful for the last several decades as evidenced by significant learning gains in studies covering a wide range of domains (e.g., Cohen, Kulik, & Kulik, 1982; Graesser, McNamara, & VanLehn, 2005; Merrill, Reiser, Ranney, & Trafletton, 1992). However, one potential weakness of long-term ITSs is that while novel to students at first, they can become repetitive over time. This facet is a particular problem when the targeted skill or knowledge requires extended practice to reach sufficient mastery or depth of understanding.

An increasing number of long-term tutoring systems focus on prolonged skill acquisition across multiple interactions, and several of these have been integrated and evaluated within ecological settings (Jackson, Boonthum, & McNamara, 2010; Johnson & Valente, 2008; Koedinger & Corbett, 2006; Meyer & Wijekumar, 2011). Due to the extended time span of these interactions, students can sometimes become disengaged and bored while using some systems (e.g., Arroyo et al., 2007; Baker, D'Mello, Rodrigo, & Graesser, 2010; Bell & McNamara, 2007). When the learning

process is expected to require multiple days, weeks, or months, designing the environment such that it induces the learner to persist should be paramount among the design objectives. If students do not remain engaged and persist within a training environment, attaining a long-term learning objective is nearly impossible. Furthermore, for those students who continue to interact despite lack of interest, boredom may trigger a vicious cycle that prevents them from actively re-engaging in constructive learning processes (Baker, Corbett, & Koedinger, 2004; D'Mello, Taylor, & Graesser, 2007).

Educational systems that require a longer training commitment may benefit from design features that enhance student engagement after any novelty effects have dissipated. Nonetheless, there is more to learning than interest and engagement. Sacrificing essential pedagogical aspects of an educational environment to increase interest is not likely to be successful. As these constraints have become more evident, system designers have begun to carefully incorporate educational games and game-based elements to help capture students' interest and promote active participation within learning environments (McNamara, Jackson, & Graesser, 2010).

Game-Based Learning

It is intuitively clear that games are a potentially strong motivating factor for students (Gee, 2003; Steinkuehler, 2006). A natural, intrinsic interest in the domain content of the system is, of course, the preferred method of obtaining involvement, but unfortunately not all learners share interests. While the content itself plays an important role for determining interest, perhaps the framing of this content (e.g., incorporating it within a game) is even more crucial. Thus, a game itself can be used as a catalyst to promote motivation and sustain the interest of students.

The increased focus on games in education may also be partially due to the alignment between aspects of game design and the goals

This article was published Online First September 9, 2013.

G. Tanner Jackson and Danielle S. McNamara, Learning Sciences Institute, Arizona State University.

This research was supported in part by the Institute for Educational Sciences (R305G040046) and the National Science Foundation (REC0241144; IIS-0735682). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these organizations.

Correspondence concerning this article should be addressed to G. Tanner Jackson, Learning Sciences Institute, Arizona State University, P.O. Box 872111, Tempe, AZ 85287-2111. E-mail: TannerJackson@asu.edu

of educational environments. This is not simply a grafting of two successful but incompatible technologies; research suggests that these technologies have a common theoretical foundation and that the sum is greater than the parts (e.g., Laird & van Lent, 2000; Van Eck, 2006). Specifically, an essential overarching benefit to games is that they, similar to tutoring systems, provide the opportunity for adaptive, individualized interactions. The notion behind these highly interactive educational games is to involve learners, giving them opportunities to perform, experience outcomes, and reflect on the targeted tasks such that these actions are integrated within a meaningful context (Barab et al., 2010).

Games often improve engagement and motivation by employing features similar to those found within successful tutoring systems. For example, one of the many motivating factors of games is the individual and personalized nature of the interactions that adapt to the skills and actions of the player (Gee, 2005; Malone & Lepper, 1987; Rieber, 1996). To accomplish this goal, an educational game must be able to identify the ability level of the learner and adjust itself accordingly (Conati, 2002; Rieber, 1996; Shute & Towle, 2003). As such, the game may require demonstration of more advanced skills or knowledge from a learner progressing successfully through the game or lessen the requirements for a learner progressing poorly. Additionally, the rapid feedback within educational games can help learners to better regulate their progress and activities. Indeed, the role of feedback in any learning environment can lend a stronghold on engagement (Anderson, Corbett, Koedinger, & Pelletier, 1995; Corbett & Anderson, 1990; Foltz, Gilliam, & Kendall, 2000). By leveraging these features to increase engagement and motivation, these games are highly compatible with the sophisticated pedagogy implemented within most ITSs.

Another aspect of games that maps onto pedagogical goals is the notion of challenge (i.e., task difficulty; Gredler, 2004; Rieber, 1996). Games that are easily won require little effort from learners. On the other hand, games that are too difficult can result in lowered interest because learners are unable to accomplish goals. Vygotsky (1978) posited that learning is most effective when the material is slightly more advanced than the learner. With respect to game challenge, the same hypothesis could apply. A game that is slightly more challenging than the learner's skill and knowledge may sustain interest and motivation by providing accomplishment while maintaining effort (Gee, 2003). Indeed, self-efficacy and interest in games have been found to be highly correlated (Zimmerman & Kitsantas, 1997). Ratings of higher self-efficacy during game play coincide with higher preferences for one game over others. Thus, accomplishment by the players over consistent challenges should raise their self-efficacy, overall enjoyment, and motivation to perform the task.

Motivation and Mastery in Educational Games

Ample research shows that learning (and mastery) is more than just a cognitive process (du Boulay, 2011); learning is as much a motivational and affective task as it is a demonstration of mental ability. Research also suggests that there is an indirect link between motivation and learning (Garris, Ahlers, & Driskell, 2002); namely, motivation influences the learning processes in which students engage. And, these processes subsequently affect learning outcomes.

Motivation is a multidimensional construct that subsumes a number of component factors, such as interest, enjoyment, expectancies, and values. For the current work, *motivation* generally refers to students' desire to perform a task and willingness to expend effort on that activity (Garris et al., 2002; Pintrich & Schrauben, 1992; Wolters, 1998). This broad conceptualization of motivation encompasses previous research examining both intrinsic and extrinsic factors related to interest, engagement, enjoyment, and self-efficacy. This prior work has indicated that enhancing these aspects of motivation positively impacts learning (Alexander, Murphy, Woods, Duhon, & Parker, 1997; Bandura, 2000; Pajares, 1996; Pintrich, 2000; Young et al., 2012; Zimmerman & Schunk, 2001). Other research has demonstrated that various *mechanisms* common to games, such as feedback, incentives, task difficulty, and control, can have a significant impact on these motivational constructs and, hence, may ultimately affect learning (Conati, 2002; Corbett & Anderson, 2001; Cordova & Lepper, 1996; Graesser, Chipman, Leeming, & Biedenbach, 2009; Malone & Lepper, 1987; Moreno & Mayer, 2005; Shute, 2008).

Many games leverage these mechanisms and other features as part of a core game design. No individual feature is required within a game, and some game elements may even be unnecessary, ineffective, or distracting in the short term, but they may also have the potential to increase interest, enjoyment, and engagement in the long term. Previous research has also suggested that the affective benefits from games may increase as the number of incorporated game-based features increases (Cordova & Lepper, 1996; Papastergiou, 2009). Therefore, some researchers have assumed that combining several game features together will provide students with a more enjoyable interaction (Asgari & Kaufman, 2004; McNamara et al., 2010).

Unfortunately, despite the increase in research related to educational games, there remains a dearth of research in which the effectiveness of these new gaming environments have been directly compared with their natural counterpart, traditional intelligent tutoring environments (O'Neil & Fisher, 2004; O'Neil, Wainess, & Baker, 2005). Two recent studies have been conducted in which researchers have directly investigated the effectiveness and benefits of educational game components compared with an ITS. The first study by Jackson, Dempsey, and McNamara (2012) was a 90-min experiment to compare the short-term practice effects of a traditional ITS environment with a game-based counterpart. They found that participants who had interacted with game-based practice rated it as significantly more engaging than students within the traditional ITS. By contrast, students who interacted with the traditional ITS outperformed students who practiced using the game environment.

A second smaller study was conducted over a longer time span (six separate sessions) to investigate a combined system that allowed users to continually choose between practicing with an ITS or a game-based system (Jackson, Dempsey, Graesser, & McNamara, 2011). Participants in this study completed a 2-hr introductory training session before entering the practice environment where they could choose between systems (for the remaining 4–5 hr across sessions). Focusing on the results comparing the same two systems from Jackson et al. (2012), there were no advantages for the traditional ITS in this longer term study in terms of performance (comparing within-subjects). The students performed equally well within both systems. In addition, although

there were trends showing improved enjoyment for the game-based system over the ITS, the difference was not statistically significant. The findings from these studies helped to motivate the current study, provide support for differing hypotheses (discussed in more detail later), and suggest that the current study is needed to further explore the complex interplay between games and learning (also see Harris, 2008).

The current work aims to more directly address these issues in game-based learning by comparing the outcomes from two similar long-term skill acquisition systems: a traditional ITS (iSTART) and an educational game (iSTART-ME).

iSTART and iSTART-ME

The Interactive Strategy Training for Active Reading and Thinking-Motivationally Enhanced (iSTART-ME) tutor is a newly developed game-based learning environment built on top of an existing tutoring system (iSTART). iSTART provides young adolescents to college-age students with comprehension strategy training to better understand and learn from challenging science texts (McNamara, Levinstein, & Boonthum, 2004; McNamara, O'Reilly, Best, & Ozuru, 2006). In iSTART, pedagogical agents instruct trainees in the use of self-explanation and other active reading comprehension strategies to explain the meaning of science text while they read. The training was motivated by empirical findings showing that students who self-explain text are more successful at solving problems, more likely to generate inferences, construct more coherent mental models, and develop a deeper understanding of the concepts covered in text (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994; McNamara, 2004).

iSTART Modules

Strategy instruction occurs in three stages with each stage requiring increased interaction on the part of the learner. During the Introduction Module of iSTART, a trio of animated characters introduces students to the concept of self-explanation and associated reading strategies by providing information, posing questions, and discussing examples. In the second phase, called the Demonstration Module, two agents demonstrate the use of self-explanation using a science text, and the trainee identifies the strategies being used by the agents. During this module, the teacher character (Merlin) asks the trainee to indicate which strategies the student agent (Genie) employed in producing his self-explanation. Finally, Merlin gives Genie feedback on the quality of his self-explanation.

In the third phase (Practice), Merlin coaches and provides feedback while the trainee practices self-explanation using the repertoire of reading strategies. The goal is to help the trainee acquire the skills necessary to integrate prior text and prior knowledge with the current sentence content. For each sentence, Merlin reads the sentence, asks the trainee to explain it by typing a self-explanation, and provides feedback on the quality of the explanation.

The iSTART assessment algorithm drives the feedback provided by Merlin. The algorithm output is coded as a 0, 1, 2, or 3. An assessment of 0 indicates that the self-explanation was either too short or contained mostly irrelevant information. An iSTART score of 1 is associated with a self-explanation that primarily relates only to the target sentence itself (sentence-based). A 2 means that the student's self-explanation incorporated some aspect of the text beyond the target sentence (text-based). If a self-explanation earns a 3, then it is interpreted to have incorporated information at a global level and may include outside information

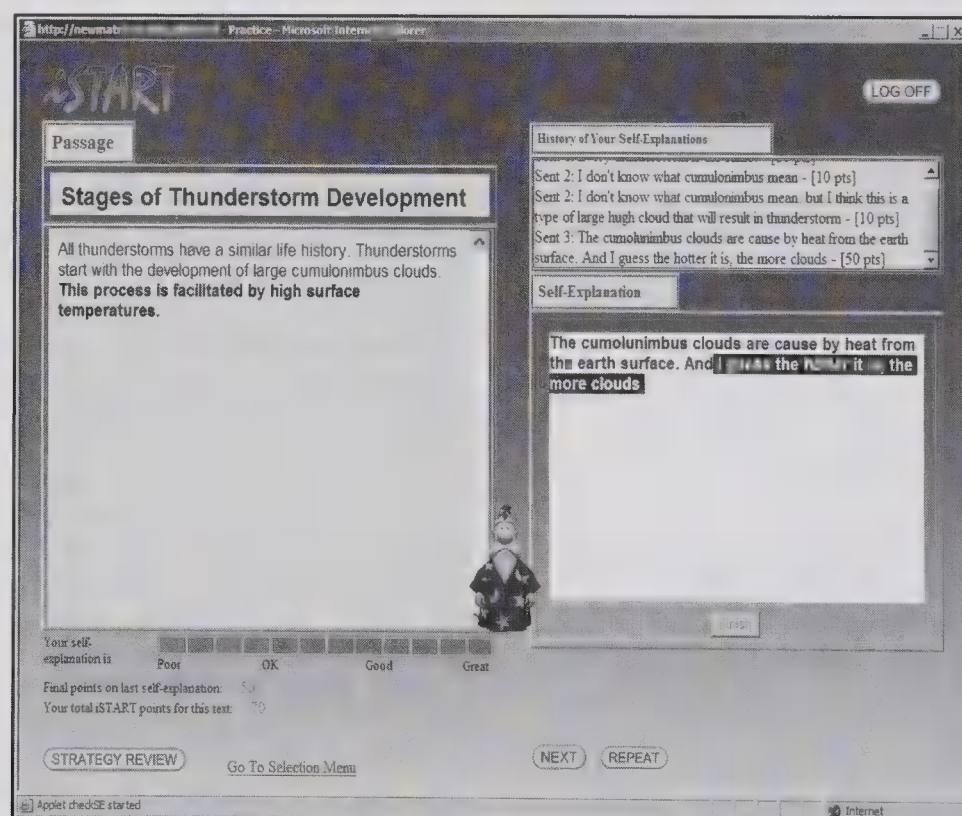


Figure 1. Screenshot of Coached Practice in iSTART.

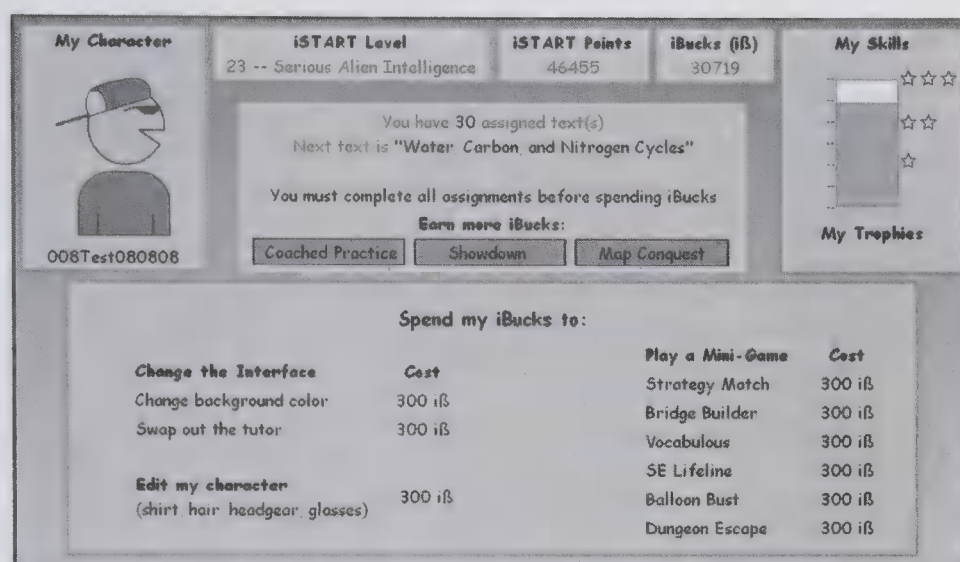


Figure 2. Screenshot of iSTART-ME selection menu.

or refer to an overall theme across the whole text (global-based). This algorithm has demonstrated performance comparable to that of humans and provides a general indication of the cognitive processing required to generate a self-explanation (Jackson, Guess, & McNamara, 2010).

Within iSTART, there are two types of practice modules. The first practice module is situated within the core context of iSTART (initial 2-hr training) and includes two texts. The second practice module is a form of extended practice, which operates in the same manner as the regular practice module. This extended practice phase (called Coached Practice—see Figure 1 for a screenshot) is designed to provide a long-term learning environment that can span weeks or months. Research has shown that this extended practice increases students' performance over time (Jackson, Boonthum, et al., 2010). However, one unfortunate side effect of this long-term interaction is that students often become disengaged and uninterested in using the system (Bell & McNamara, 2007).

iSTART-ME

Previous research with iSTART pointed to the need for students to persist within the system across several days of training. Therefore, changes were implemented within the system to combat the problem of disengagement over time. The extended practice module of iSTART was redesigned and situated within a game-based environment called iSTART-ME (Motivationally Enhanced). This game-based environment was built directly on top of the existing iSTART system. The main goal of the iSTART-ME project was to implement several of the game-based principles and mechanisms that were expected to support effective learning, increase motivation, and sustain engagement throughout a long-term interaction with an established ITS. The project attempted to implement and potentially manipulate these motivational constructs via game-based features that map onto one of five interaction mechanisms: feedback, incentives, task difficulty, control, and environment (see McNamara et al., 2010, for more details on the mechanisms).

The original ITS version of iSTART with Coached Practice automatically progresses students from one text to another with no intervening tasks. The new version of iSTART-ME is situated within a cohesive meta-game and point-based economy that the

user can control through a selection menu (see Figure 2 for screenshot). This new selection menu provides students with opportunities to interact with new texts (control/task difficulty), earn points and trophies (feedback/incentives), advance through levels (feedback/incentives), unlock new features (control/incentives/environment), purchase rewards (control/incentives/environment), personalize a character (control/incentives/environment), and play educational mini-games (control/incentives/task difficulty).

Within iSTART-ME, students earn points as they interact with texts and provide their own self-explanations. Each time a student submits a self-explanation, it is assessed by the iSTART algorithm and points are awarded based on a scoring rubric. The rubric has been designed to reward consistently good performance. So students earn more points if they repeatedly provide good self-explanations but earn fewer points if they fluctuate between good and poor performance. These points help go beyond the qualitative responses from the animated agents to provide an additional, quantifiable form of *feedback* as students learn and practice the self-explanation strategies. For example, students can easily understand that a score of 30 is better than a score of 10, but it is more difficult to gauge the relative difference between, "All right, let's keep going" and "You're starting to get the hang of this." In addition to serving as a form of feedback, earning points within iSTART-ME serves two main *incentive* purposes: advancing through levels and purchasing rewards.

As students accumulate more points, they advance through a series of levels. Each subsequent level requires an increasing number of points. Therefore, students must expend slightly more time or effort for further advancement (i.e., increasing *task difficulty* to reach a new level). Whenever students advance up a level, a new subset of features is automatically unlocked and becomes available within the interface (thus acting as an *incentive* and providing additional *control*). Each of the iSTART levels are labeled (e.g., *ultimate bookworm*, *serious strategizer*) to help provide incentive, increase interest, and serve as global indicators of progress across texts.

Points can also be used to *control* the environment by "purchasing" incentives within the system. One of the options available as a reward allows students to change aspects of the learning

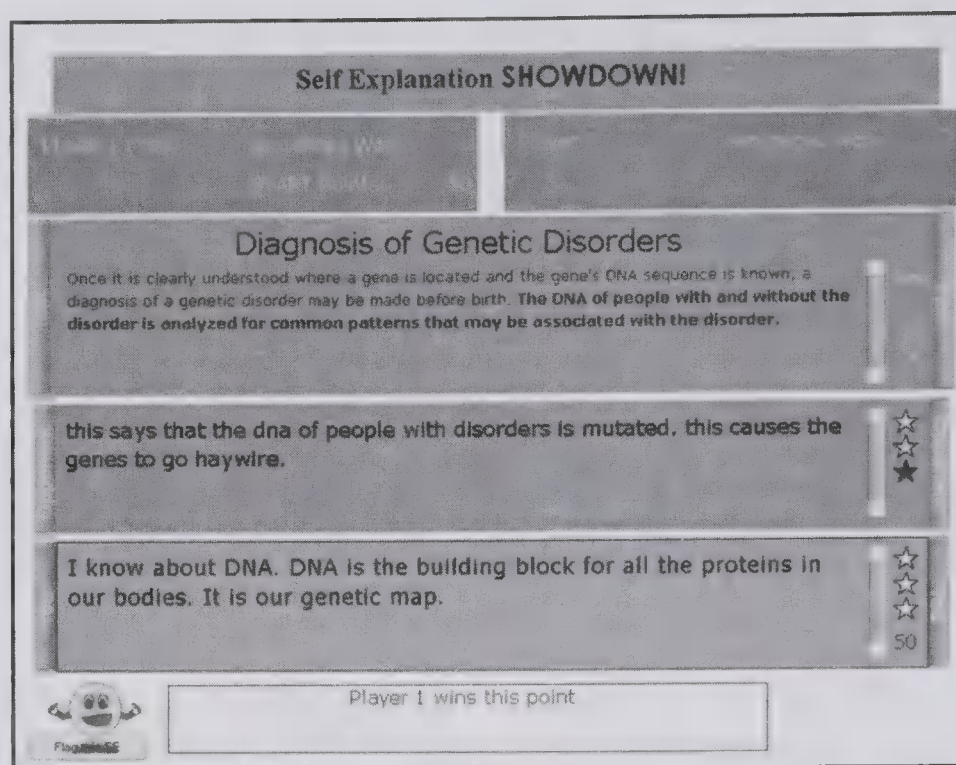


Figure 3. Screenshot of Showdown in iSTART-ME.

environment. They can spend some of their points to choose a new tutor agent, change the interface to a new color scheme, or update the appearance of their personalizable avatar. These features provide students with a substantial amount of control and personalization over their environment and have been designed as purchasable replacements, rather than continuously available options, to help reduce off-task behaviors (such as switching back and forth between agents solely to see what they all look like).

Last, a suite of eight educational mini-games have been designed and incorporated within the iSTART-ME extended practice module. Some mini-games require identification of the type of strategy use, while others may require students to generate their own self-explanations. The majority of iSTART-ME mini-games require similar cognitive processes enveloped within different combinations of gaming elements.

Showdown and Map Conquest are two methods of generative game-based practice that use the same iSTART assessment algorithm from regular practice. In Showdown (see Figure 3 for a screenshot), students compete against a computer player to win rounds by producing better self-explanations. After the learner submits a self-explanation, it is scored, the quality assessment is represented as a number of stars (0–3), and an opponent self-explanation is also presented and scored. The difficulty of the opponent self-explanation has been manipulated within previous experiments (Dempsey, 2011); however, for normal gameplay, the opponent example is chosen at random to provide a range of student modeling (i.e., good and bad examples). The self-explanation scores are compared, and the player with the most stars wins the round. The player with the most rounds at the end of the text is declared the winner. The combination of features for Showdown incorporates aspects of feedback (points, stars, rounds won), incentives (points, stars),

control (production of self-explanation), and task difficulty (opponent, text content).

Map Conquest is the other game-based method of practice where students generate their own self-explanations. Within Map Conquest, the quality of a student's self-explanation determines the number of dice that student earns (i.e., performance at the target task determines the resources available during a subsequent game task). Students place these dice on a map and use them to conquer neighboring opponent territories, which are controlled by two virtual opponents. The surface components are somewhat different from those in Showdown but were similarly designed to provide the user with feedback (points, dice), incentives (dice, map puzzle), control (map puzzle, production of self-explanation), and task difficulty (opponents, text content).

In most of the identification mini-games—for example, Balloon Bust (Figure 4)—students are presented with a target sentence and an example self-explanation. The student must decide which iSTART strategy was used in the self-explanation and then click on the corresponding balloons. There are also three other mini-games that focus on the same task of identifying strategies within example self-explanations. These other games each incorporate a new interface with a different combination of game elements, which might include fantasy, competition, and perceptual aspects (as in Balloon Bust). Though the surface features of these games can differ widely, they have been designed with very similar underlying mechanisms and can all be completed within 10–20 min. Students are allowed to select any form of practice or mini-game from the selection menu that has been unlocked (provided that they have enough points). After completion of a task, students are directed back to the main iSTART-ME selection screen.

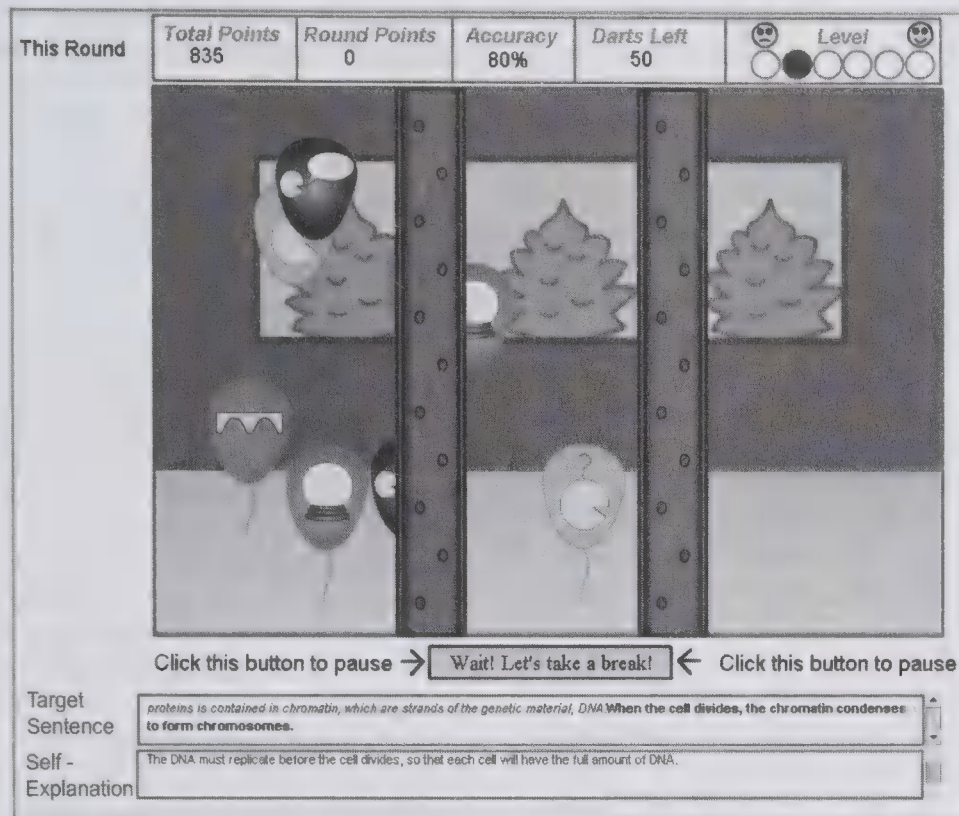


Figure 4. Screenshot of Balloon Bust in iSTART-ME.

Current Study

The current study was a multisession experiment in which the effectiveness of a game-based tutoring system (iSTART-ME) was compared with its ITS counterpart (iSTART-Regular). One possible concern with integrating games into learning systems is that they have the potential to detract from the immediate pedagogical goals and reduce learning improvements in the short term (Jackson et al., 2012; Mayer & Moreno, 2003; Paas, Renkl, & Sweller, 2003). However, across long-term training, the engagement fostered by the game environment may compensate for any distracting elements, thus allowing students to catch up in performance (Jackson, Dempsey, Graesser, & McNamera, 2011). Hence, this study was conducted to thoroughly explore the potential long-term benefits of game-based training, how it compares with training from a traditional ITS, and how various effects of motivation and learning may unfold over time.

Hypotheses

The Jackson et al. (2012) study indicated that students who received game-based training during early stages of skill acquisition exhibited decreased performance at the target task (compared with students in a traditional ITS). In contrast, the Jackson, Dempsey, et al. (2011) study showed that when students completed initial strategy training within a traditional ITS (i.e., no game features), subsequent performance during game and nongame practice methods was equivalent. Additionally, previous work with the game-based aspects in iSTART-ME has shown consistent positive effects for motivation and enjoyment (Jackson, Davis, et al., 2011; Jackson & McNamara, 2011). This combination of results leads to two hypotheses regarding the current study.

One hypothesis is that games improve motivation and enjoyment, but they may impede learning, especially initially (Adams, Mayer, MacNamara, Koenig, & Wainess, 2012; Jackson et al., 2012). In this case, we would expect the game-based environment to produce lower learning outcomes than the traditional ITS, particularly in the *initial stages of learning*. The second hypothesis is that the game-based components of iSTART-ME improve motivation and enjoyment (Cordova & Lepper, 1996; Papastergiou, 2009), and this increase in affective measures mediates learning (Alexander et al., 1997). This hypothesis suggests that students in the game-based training should see improved motivation and enjoyment over time and should see a corresponding increase in performance during the *later stages of training* (compared with the traditional ITS).

Procedure

Participants and setting. Eighty-four high school students were recruited from the general city-wide high school population in an urban environment in the mid South (51% male; 81% African American, 13% White, 6% other; average grade completed = 10th grade; average age = 15.8 years). The 11-session experiment was conducted in a research laboratory on a large university campus and involved four phases: pretest, training, posttest, and retention test.

Pretest. During the first session, students completed a pretest that included questions to collect basic demographics, prior motivation (including selected questions adapted from the Motivated Strategies for Learning Questionnaire, or MSLQ; Pintrich, Smith, Garcia, & McKeachie, 1993), and an assessment of their prior ability to self-explain (described in more detail later).

Training. At beginning of the eight training sessions, participants completed a 12-item daily survey (see Measures section for details). After the daily survey, students then interacted with their randomly assigned between-subjects condition: a game-based system (iSTART-ME, $n = 41$) or a traditional ITS (iSTART-Regular, $n = 43$). Students in the educational game condition interacted with the full game-based selection menu in iSTART-ME across eight separate sessions of at least 1 hr each. Participants in the ITS condition used the original non-game-based version of iSTART for the same amount of time (eight sessions of at least 1 hr each).

The initial training within both conditions was identical until the participants transitioned into extended practice. That is, both conditions progressed through the Introduction Module, the Demonstration Module, and then two regular practice texts within the Coached Practice environment. Students assigned to game-based training were then free to use the full selection menu (Figure 2), while the ITS students continually transitioned from one text to another within the Coached Practice environment (Figure 1).

Like many ITSs, iSTART-Regular is not completely void of mechanisms and features that are commonly used within games. For example, iSTART-Regular displayed points for each self-explanation (near bottom-left of Figure 1), included adaptive feedback from an animated agent and provided a trophy (or lack thereof) based on the performance within each text. These features (points, personalized feedback, animated characters, and trophies/badges) are commonly used in numerous types of games and systems, both virtual and physical. Table 1 provides a more thorough comparison of the two training systems in terms of the key features included in each. iSTART-ME differed from iSTART-Regular primarily in the presence of the selection menu, which allowed participants to play mini-games and modify certain aspects of the environment (e.g., swap tutors, personalize their avatar). Both systems allowed students to progress through the tutoring at their own pace, and therefore, not all students experienced the same components at the same time. This is a key characteristic of ITSs and virtually all games that adapt interactions on the basis of user decisions. Hence, some students naturally receive more or different kinds of training and practice than others.

Posttest and retention. All students completed the posttest and then a delayed retention test (completed a week after posttest). The posttest consisted of assessments similar to those from the pretest (details are discussed in the Measures section). These included measures of self-explanation ability and students' motivation during the study, along with questions pertaining to students' attitudes, perceptions, and experiences. The retention test

was used to assess the durability of students' self-explanation skills after a 1-week delay without training.

Measures

Survey and performance measures were collected during pretest, training, posttest, and retention. These included measures related to self-explanation ability as well as students' attitudes, motivation, self-efficacy, and enjoyment.

Self-explanation ability. Students' performance on self-explanation tasks was collected during pretest, training, posttest, and retention. During training, students interacted with various texts and all self-explanations were scored through the iSTART assessment algorithm and recorded into a database. During each of the three testing phases, students were presented with a new text (not included within training) and prompted to self-explain specific sentences (eight self-explanations during each test). These three texts were selected due to their similarity in terms of length (281–329 words), content difficulty (Grade Level 8–9), and linguistic features (i.e., similar scores on the five principal component scores within Coh-Metrix; Graesser, McNamara, & Kulikowich, 2011). Each self-explanation provided by the students was scored using the iSTART assessment algorithm, the performance of which has been shown to be comparable to that of human scorers (Jackson, Guess, et al., 2010). Unfortunately, due to a technical error, the three texts were not automatically counterbalanced across the testing phases. Thus, despite extensive efforts to utilize equitable texts, comparisons of self-explanations across time should be interpreted with caution and must be replicated using appropriate methodology. Nonetheless, the lack of counterbalancing should not affect any comparisons between conditions.

Attitudes, motivation, self-efficacy, and enjoyment. Survey questions were included during pretest, posttest, and daily training sessions to assess students' attitudes, motivation, self-efficacy, and enjoyment. Pretest and posttest measures included several questions adapted from the MSLQ (Pintrich et al., 1993). The questions adapted from the MSLQ were selected to address students' motivation and self-efficacy. In addition to these standardized measures, questions were included from previous research with the iSTART system (Jackson, Davis, et al., 2011; Jackson, Graesser, & McNamara, 2009; Jackson & McNamara, 2011). These additional questions were implemented within the pretest, posttest, and daily surveys and were designed to measure students' self-assessments of motivation, expectations for system interactions, current affect and mood, and overall enjoyment of the system.

Table 1
Game Mechanism and Feature Differences Between iSTART-ME and iSTART-Regular

Mechanisms	iSTART-ME	iSTART-Regular
Feedback	Points, local skill bar, verbal feedback from pedagogical agent, global skill bar, trophies, levels	Points, local skill bar, verbal feedback from pedagogical agent
Incentives	Points, trophies (reviewable), levels, swap tutor, edit theme, edit character, play mini-game	Points, trophies (viewed once after each text)
Control	Select next activity, edit environment & characters, generate self-explanations	Generate self-explanations
Task difficulty	Increased difficulty for each new level (both in menu and in games), new texts	New texts
Environment	Animated pedagogical agents, select new animated agent, edit theme, edit character, display trophy case, display performance for recent texts	Animated pedagogical agents

Table 2
Pretest and Posttest Survey Scales: Means (Standard Deviations)

Survey scales	iSTART-ME	iSTART-Regular	<i>F</i> (1, 82)	<i>p</i>
Pretest survey measures				
Expected Enjoyment and Motivation to Participate (5 items)	4.73 (0.66)	4.69 (0.73)	0.07	.799
Achievement Motivation and Learning Values ^a (7 items)	5.31 (0.98)	5.31 (0.86)	0.00	.996
Self-Efficacy (3 items)	5.94 (1.12)	5.95 (0.91)	0.00	.962
Competitiveness (2 items)	4.91 (1.08)	4.84 (1.21)	0.10	.758
Posttest survey measures				
Enjoyment and Motivation (6 items)	4.55 (1.09)	3.83 (1.20)	8.28	.005
Perceived Learning, Effort, and Values ^a (9 items)	4.81 (1.56)	4.53 (1.61)	0.64	.425
Ease of Use (4 items)	3.32 (1.04)	3.31 (1.20)	0.00	.951

^a Questions adapted from the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich, Smith, Garcia, & McKeachie, 1993).

The daily surveys used within the current study have been used in previous research with the game-based version of iSTART-ME (Jackson, Davis, et al., 2011; Jackson & McNamara, 2011). These surveys were designed to assess students' moods, attitudes, and perceptions across the time frame of an experiment without being an invasive measurement during system interactions. These surveys were administered at the beginning of each training session and specifically addressed students' experiences from the previous session (overall impression, enjoyment, boredom, frustration, interface problems, feeling of learning, feeling of improvement) and assessed their current attitudes and feelings (current mood, anticipation about participating, level of motivation, intention to perform well, desire to do better than others).

Results

Training Sessions

As mentioned previously, both systems allowed students to progress through training at their own speed. Despite the adaptivity and self-paced interactions, students' prior ability was not related to the amount of practice students received during training. More specifically, self-explanation ability at pretest was not related to the number of practice texts that students completed ($r = .079$, $p = .45$). Therefore, initial ability levels were not related to the amount of extended practice that students received, and most students experienced the training components at approximately the same time. The vast majority of students completed the two regular practice texts and transitioned into the extended practice during the first ($n = 6$) or second ($n = 72$) session, while some students did not reach the extended practice section until the third ($n = 5$) or fourth ($n = 1$) session. Ultimately, all students completed the training modules and subsequently interacted with their randomly assigned training condition for the remainder of the study.

Attitudes, Motivation, Self-Efficacy, and Enjoyment

User experience measures from pretest questions, daily surveys, and posttest questions were analyzed to explore students' attitudes, perceptions, and experiences within the two training systems. Analyses on the pretest survey questions indicate that there were no significant differences between conditions on questions prior to the start of training that related to enjoyment, motivation, self-efficacy, or competitiveness (see Table 2).

The posttest survey included several questions related to enjoyment, perceived learning, and usability within the system (see Table 2 for descriptive and analysis of variance [ANOVA] results). A posttest enjoyment and motivation composite score was created by averaging across six separate questions. An ANOVA on the enjoyment and motivation composite score yielded a significant effect of condition, $F(1, 82) = 8.28$, $p = .005$, mean square error (MSE) = 1.15, Cohen's $d = 0.628$. These results indicate that the game-based environment was rated as a significantly more positive experience than the traditional ITS. Additionally, a composite scale that assessed students' perceived learning, effort, and values for the target system and materials found no significant differences between the game and nongame system. Likewise, a four-question scale that assessed system ease of use and interface confusion revealed no significant differences between conditions. These results suggest that the game-based selection menu system was more enjoyable and motivating, but just as valuable and easy to use as the ITS.

Daily surveys were administered to assess students' reports of their previous-session experiences and current-session expectations. Questions related to similar concepts were combined into several composite scores (i.e., enjoyment during the previous session, improvements in self-efficacy, and motivation for the current session). A composite score was created for enjoyment during the previous session by combining scores from the following three questions: "My most recent session was . . . (*very bad* = 1, *very good* = 6)," "I enjoyed my most recent session . . . (*not at all* = 1, *very much* = 6)," and "I was bored during my most recent session . . . (reversed scored; *all the time* = 1, *never* = 6)." A mixed-factor ANOVA on this composite score indicated that there was a significant main effect for condition, such that students in the game-based condition rated their session experiences more favorably ($M = 4.89$, standard error [SE] = 0.159) than did students in the ITS condition ($M = 4.07$, $SE = 0.151$), $F(1, 76) = 13.92$, $p < .001$, $MSE = 7.51$. There was also a significant linear interaction between session and condition, $F(1, 76) = 3.266$, $p = .004$, $MSE = 0.606$ (see Figure 5).¹ Pairwise comparisons using Bonferroni adjustments for multiple tests confirmed that enjoy-

¹ The mixed-factor ANOVA results presented here are based on the linear equation contrasts for the interaction. The overall within-subject interaction effects for this mixed-factor ANOVA (including Huynh-Feldt corrections due to significant sphericity and a large Greenhouse-Geisser $\epsilon > .75$) were also significant, $F(5.99, 455.26) = 3.27$, $p = .004$.

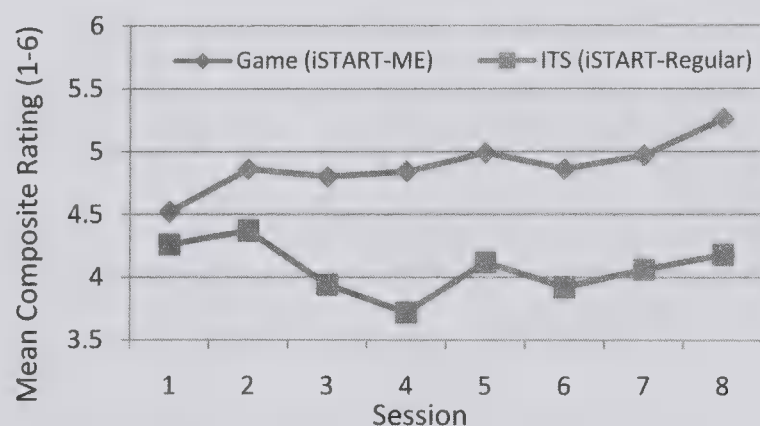


Figure 5. Composite means for enjoyment questions about the previous session. ITS = intelligent tutoring system.

ment tended to increase across sessions for students interacting with the game-based system and decrease for those in the ITS condition. Specifically, for students in the iSTART-ME condition, overall enjoyment at Sessions 5 ($t = 4.18$, $p_{\text{adjusted}} < .01$) and 8 ($t = 5.03$, $p_{\text{adjusted}} < .001$) were significantly higher than at Session 1 (with the middle sessions being roughly equivalent). By contrast, students in the ITS condition provided their highest ratings during the first two sessions, with enjoyment at Session 2 being significantly higher than at Session 4 ($t = 3.40$, $p_{\text{adjusted}} < .05$).

Similarly, a composite score was created for the daily survey questions related to students' improvements in self-efficacy. This score combined student ratings for two questions about the previous session, "I felt like I learned the material . . . (not at all = 1, very much = 6)" and "I feel like my reading skills improved . . . (not at all = 1, very much = 6)." A mixed-factor ANOVA on the self-efficacy composite score did not indicate a significant main effect for condition, $F(1, 76) = 2.50$, $p = .118$, $MSE = 7.43$, but did reveal a significant linear interaction between session and condition, $F(1, 76) = 2.91$, $p = .015$, $MSE = 0.673$ (see Figure 6).² This interaction reflects the finding that students' reported self-efficacy increased across sessions if they had interacted with the game-based version of training and decreased if they interacted with the traditional ITS. Specifically, pairwise comparisons (using Bonferroni adjustments) showed that iSTART-ME students provided their highest self-efficacy rating in the final session (Session 8 was marginally higher than Session 4, $t = 3.15$, $p_{\text{adjusted}} = .09$).

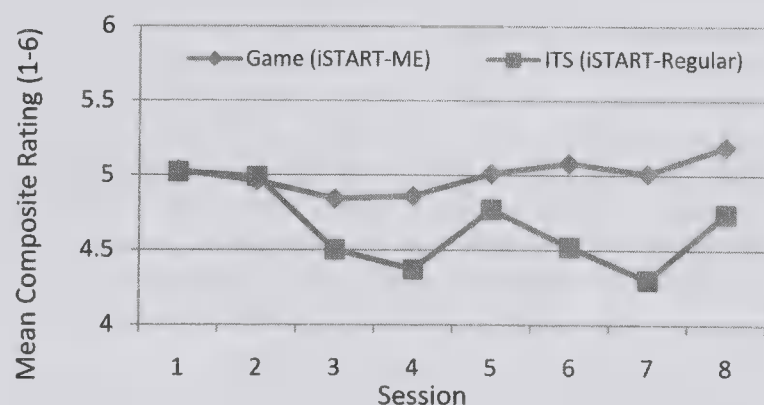


Figure 6. Composite means for self-efficacy daily survey questions. ITS = intelligent tutoring system.

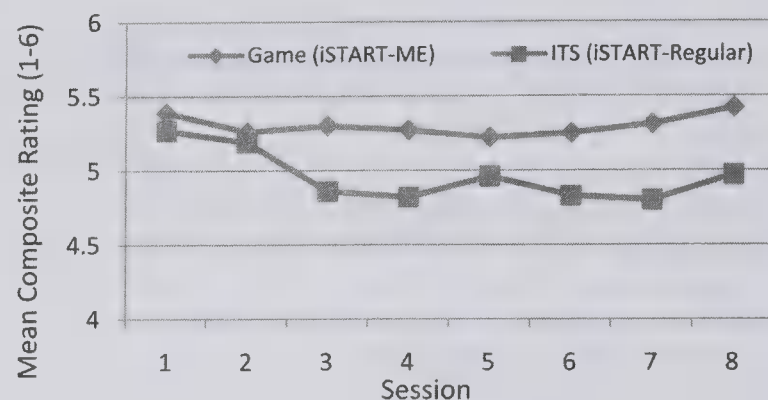


Figure 7. Composite scores for questions about students' motivation to participate in the current session. ITS = intelligent tutoring system.

In contrast, iSTART-Regular students provided their highest self-efficacy ratings in the first two sessions (Session 1 was significantly higher than Session 7, $t = 3.35$, $p_{\text{adjusted}} < .05$; and Session 2 was significantly higher than both Session 4, $t = 3.57$, $p_{\text{adjusted}} < .05$, and Session 7, $t = 3.52$, $p_{\text{adjusted}} < .05$).

Finally, a composite score was created for the daily survey questions that pertained to motivation to participate in current session: "My mood right now is . . . (very negative = 1, very positive = 6)," "I am looking forward to participating in today's session . . . (not at all = 1, very much = 6)," "I am motivated to participate in today's session . . . (not at all = 1, very much = 6)," and "I plan to do my best during today's session . . . (not at all = 1, very much = 6)." A mixed-factor ANOVA yielded a marginal main effect of condition, $F(1, 75) = 3.05$, $p = .085$, $MSE = 5.82$, indicating that students in the game-based condition ($M = 5.30$, $SE = 0.142$) tended to be more motivated to participate than students interacting with the ITS ($M = 4.96$, $SE = 0.133$). This mixed-factor ANOVA also revealed a significant linear interaction between session and condition, $F(1, 75) = 4.95$, $p = .029$, $MSE = 0.410$ (see Figure 7),³ reflecting the finding that students' motivation to participate in the current session remained stable for those in the iSTART-ME condition but declined in the iSTART-Regular condition. Pairwise comparisons (using Bonferroni adjustments) confirmed that students' ratings for today's session within the game-based system were not significantly different across sessions ($p_{\text{adjusted}} > .05$), and students within the ITS provided marginally higher ratings in Sessions 1 and 2, compared with Sessions 4 ($t_{\text{Session1}} = 3.24$, $t_{\text{Session2}} = 3.26$, $p_{\text{adjusted}} < .10$) and 7 ($t_{\text{Session2}} = 3.15$, $p_{\text{adjusted}} < .10$).

These results collectively indicate that students provided equivalent ratings in the two conditions for the first two sessions (when training was the most similar), but after the game-based aspects

² The mixed-factor ANOVA results presented here are based on the linear equation contrasts for the interaction. The overall within-subject interaction effects for this mixed-factor ANOVA (including Greenhouse-Geisser corrections due to significant sphericity and a moderate Greenhouse-Geisser $\epsilon < .70$) were also significant, $F(4.75, 360.84) = 2.91$, $p = .015$.

³ The mixed-factor ANOVA results presented here are based on the linear equation contrasts for the interaction. The overall within-subject interaction effects for this mixed-factor ANOVA (including Huynh-Feldt corrections due to significant sphericity and a large Greenhouse-Geisser $\epsilon > .80$) were also significant, $F(6.22, 466.40) = 2.09$, $p = .050$.

were made available, students interacting with the educational games provided more positive ratings than did students interacting with the ITS. In sum, the combined evidence from the daily surveys and posttest questions indicates that students preferred to interact with the game-based system more so than the traditional tutoring system.

Learning Outcomes

Analyses were conducted on the self-explanation scores from the pretest, posttest, and retention test. All self-explanations were scored using the iSTART assessment algorithm which has high correspondence to human scores ($\kappa = .646$; Jackson, Guess, et al., 2010; McNamara, Boonthum, Levinstein, & Millis, 2007). As shown in Figure 8, self-explanation quality improved from pretest to posttest for students in both conditions, and this increase in performance was maintained in a delayed retention test 1 week later, but there was no benefit for either condition. Specifically, a mixed-factor ANOVA confirmed a main effect of test, $F(1, 82) = 22.67, p < .001, MSE = 0.20$, reflecting the finding that self-explanation quality scores did not differ from posttest to retention test ($t < 1$), but both the posttest ($t = 7.19, p < .001$) and retention test ($t = 7.77, p < .001$) were significantly higher than the pretest (see Figure 8). There was no effect of condition, $F(1, 82) = 1.61, p = .21, MSE = 0.71$, and no interaction between condition and test, $F(1, 82) = 0.48, p = .49, MSE = 0.17$.

One of the limitations of this study is that the self-explanation texts were not counterbalanced among the pretest, posttest and retention test phases. Therefore, the pretest to posttest improvement in self-explanation ability is conflated with a potential text effect. However, combining these findings with the improvement of self-explanation ability during training lends support to students' improvement between testing phases. Additionally, the lack of counterbalancing does not affect the comparisons between conditions at each phase. Thus, the equivalent performance between conditions at each testing phase is not confounded by text.

We also conducted analyses to examine self-explanation performance comparing conditions during extended training. The first training sessions included the complete Introduction and Demonstration modules, along with the first two texts in regular practice (initial ~2 hr training). On average, students began interacting with the two different extended practice modules during the sec-

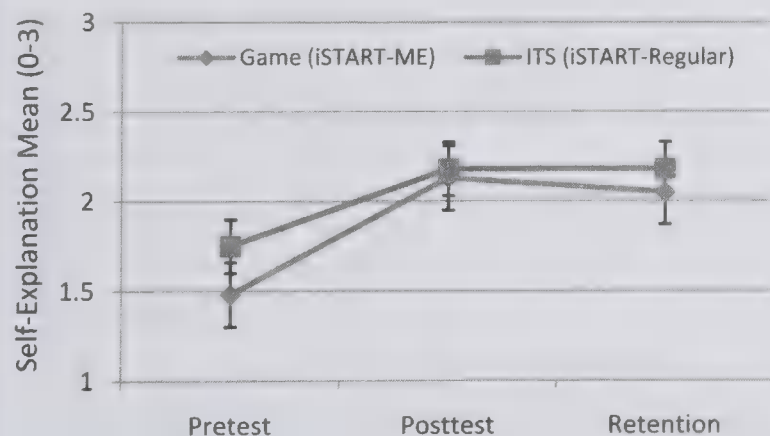


Figure 8. Self-explanation performance during testing phases (means and standard errors). ITS = intelligent tutoring system.

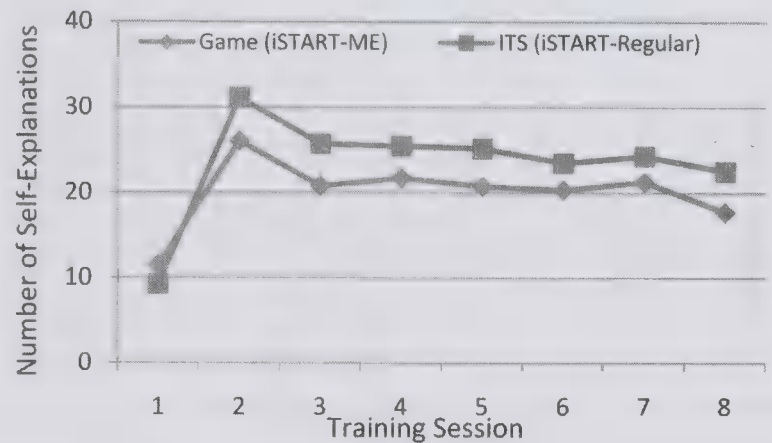


Figure 9. Frequency of self-explanations across sessions. ITS = intelligent tutoring system.

ond session (i.e., students started using either only coached practice or the full selection menu during the second session). A mixed-factor ANOVA on the frequency of self-explanations yielded significant main effects for both session, $F(1, 67) = 20.15, p < .01, MSE = 35.71$, and condition, $F(1, 67) = 4.03, p < .05, MSE = 384.86$, but revealed a nonsignificant interaction between session and condition, $F(1, 67) = 1.13, p = .29, MSE = 46.18$ (see Figure 9 for mean frequencies across days).⁴ Students who interacted with the ITS produced more self-explanations ($M = 23.39, SE = 1.14$) during extended practice than did students within the game-based training ($M = 20.03, SE = 1.23$), and the number of self-explanations tended to be highest in Session 2. Pairwise comparisons (using Bonferroni adjustments) indicate that students across conditions generated an equivalent number of self-explanations during regular practice (i.e., Session 1) and that the frequency of self-explanations during the first session was significantly lower than in all other sessions (all $t_s > 6.60$, all $p_{s_{adjusted}} < .05$). For training that took place within extended practice (i.e., Sessions 2–8), participants within the ITS produced significantly more self-explanations than students using the game-based system during Sessions 3 and 5 ($t_{Session3} = 2.16, p_{adjusted} = .035$; $t_{Session5} = 2.04, p_{adjusted} = .046$) and were marginally higher for Sessions 2 and 8 ($t_{Session2} = 1.97, p_{adjusted} = .053$; $t_{Session8} = 1.91, p_{adjusted} = .060$).

Further analyses examined students' self-explanation quality as computed by the iSTART assessment algorithm across the eight training sessions (see Figure 10). The two main hypotheses in the current study predicted opposite slopes for game-based performance during the initial and later training sessions. Specifically, the first hypothesis predicted a negative slope for game-based performance during the early sessions, while the second hypothesis predicted a positive slope for game-based performance in later sessions (this predicted decrease followed by an increase was tested through a quadratic contrast). A mixed-factors ANOVA did not indicate a significant main effect for condition, $F(1, 67) =$

⁴ The mixed-factor ANOVA results presented here are based on the linear equation contrasts for the interaction. The overall within-subject interaction effects for this mixed-factor ANOVA (including Greenhouse–Geisser corrections due to significant sphericity and a moderate Greenhouse–Geisser $\epsilon < .70$) were marginally significant, $F(4.62, 309.69) = 1.994, p = .085$.

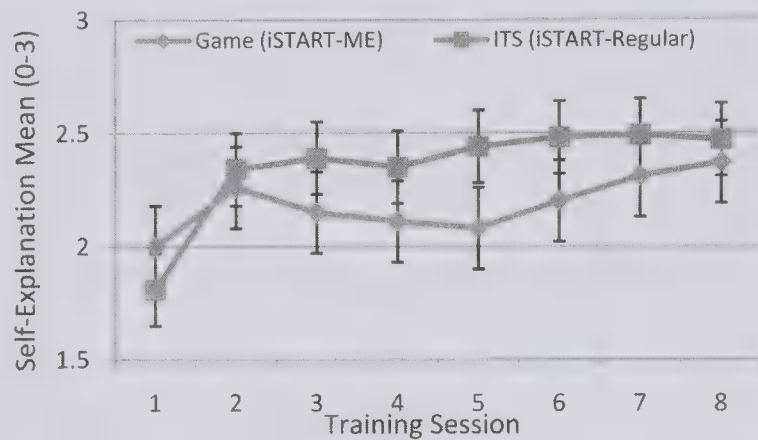


Figure 10. Self-explanation performance during training (means and standard error). ITS = intelligent tutoring system.

2.89, $p > .05$, $MSE = 1.27$; however there was a significant quadratic interaction between session and condition, $F(1, 67) = 22.91$, $p < .001$, $MSE = 0.11$.⁵ Students' self-explanation quality in the iSTART-ME condition tended to decrease during the initial interactions with the game-based selection menu and educational games (Sessions 3, 4, and 5), and Bonferroni-adjusted pairwise comparisons indicated that scores were significantly different between the two training conditions for Sessions 3 ($t = -2.05$, $p_{\text{adjusted}} < .05$), 5 ($t = -2.77$, $p_{\text{adjusted}} < .01$), and 6 ($t = -2.21$, $p_{\text{adjusted}} < .05$). These results are partially attributable to the reduction in direct feedback from Merlin in Coached Practice. These trends may also be partially due to the additional cognitive tasks involved with learning the menu itself, its features, and various game dynamics, in addition to the targeted self-explanation strategies. Indeed, the analyses related to Figure 9 demonstrate that students within the game-based system produced fewer self-explanations and thus practiced less on the target task. Despite these extra features and time spent off-task (i.e., not practicing), the students within the game-based system were able to compensate for the initial deficit over time and ultimately rose to match the performance of the ITS participants. It is important to note that the students within the game-based training were more motivated to participate, enjoyed interacting with the system more, and had larger improvements in self-efficacy than those students in the ITS condition, which would be a crucial factor in a real-life situation such as a classroom or practicing at home.

These results concur with findings in two studies conducted by Jackson, Dempsey, et al. (2011, 2012), collectively suggesting that game elements have the potential to detract from learning during initial skill acquisition. However, game environments can provide a more positive experience over time. Thus, the game-based system investigated in this study appears to strike an appropriate balance between both learning and enjoyment, improving on the imbalance previously encountered within a traditional tutoring system. This finding is especially encouraging for strategy-based tutors that require long-term interactions for students to develop skill mastery.

Discussion

The goal of tutoring systems and educational games is to produce effective and enjoyable learning experiences. However, if

students do not enjoy the experience, they are likely to cease or avoid further interactions, which is particularly detrimental to systems that require skill development over longer periods of training. These long-term skill acquisition systems must be designed to foster significant increases in mastery development, but they must also be enjoyable to use. In the case of strategy tutors, these systems must not only teach the strategies themselves but provide an effective, motivating practice environment where students can apply this training and sufficiently develop the target skills into more automatic and stable processes. For example, previous research with iSTART has illustrated the need for prolonged training of at least 5 days with the system, such that students (specifically those with low prior abilities) have sufficient opportunities to apply and master the target skills (Jackson, Boonthum, et al., 2010). Thus, a game-based version of the system (iSTART-ME) was designed to maintain higher levels of student motivation and engagement over an extended practice period by incorporating and leveraging mechanisms that positively influence affect (Conati, 2002; Corbett & Anderson, 2001; Cordova & Lepper, 1996; Graesser et al., 2009; Malone & Lepper, 1987; Moreno & Mayer, 2005; Shute, 2008). The current work focused on evaluating the global impacts of this game-based learning environment and comparing it to a traditional ITS. Additionally, in this study we investigated the specific time-based effects of these systems on both motivational and learning outcomes.

Within the current study, the game-based version of training was preferred significantly more than the traditional tutoring system. The results from the posttest survey indicate that students perceived both systems to be equally helpful and easy to use but that the game-based system was significantly more motivating and enjoyable (Table 2). Likewise, results from the daily surveys (Figures 5–7) illustrate that students who interacted with the game-based system tended to improve in their perception of the system across sessions, have improved self-efficacy (compared with those interacting with the ITS), and slowly increase (or at least maintain) motivation for future interactions. In contrast, daily ratings by students who interacted with the traditional tutoring system decreased in enjoyment, motivation, self-efficacy, and desire for future interactions. The game components present within iSTART-ME seem to be activating related motivational constructs that remain effective across time. These trends are also fairly gradual, indicating that changes may occur in smaller increments and slowly build up with more iterative interactions (possibly suggesting a cycle of affective improvement across time).

The results for self-explanation performance (the targeted skill) provide a more complex message. The self-explanation frequencies and means (see Figures 9 and 10) help to provide significant insight into the learning trajectories comparing game-based and traditional tutoring systems. Students within game-based training generated significantly fewer self-explanations than students using the ITS. This difference is likely due to time spent with the additional nongenerative activities available within the game-based selection menu (i.e., mini-games, personalizing their avatar,

⁵ The mixed-factor ANOVA results presented here are based on the contrasts for a quadratic interaction. The overall within-subject interaction effects for this mixed-factor ANOVA (including Greenhouse–Geisser corrections due to significant sphericity and a moderate Greenhouse–Geisser $\epsilon < .75$) were also significant, $F(5.00, 335.18) = 4.23$, $p = .001$.

selecting a character, changing the interface colors, and so on). Despite the increased number of self-explanations for the ITS condition, both training systems produced equivalent self-explanation performance at the posttest and delayed retention test (Figure 8).

Based on the analyses for Figure 10, it appears that the traditional ITS system showed a predominantly positive relation between the amount of training and performance. This trend was expected, based on past research substantiating the positive benefits of iSTART training over time (Jackson, Boonthum, et al., 2010; McNamara, O'Reilly, Rowe, Boonthum, & Levinstein, 2007). The trajectories for self-explanation quality within the game-based system allow us to address the primary hypotheses. Our two main hypotheses regarded the potential benefits or hindrances from the game-based version of training. Specifically, the first hypothesis predicted that the addition of game-based features may detract from the learning objectives, such that students should exhibit a decrease in performance during the initial stages of training. Indeed, the decrease in performance during Sessions 3 through 5 (Figure 10) suggest that the game-based features may initially detract or interfere with students' ability to apply the target strategies (possibly due to competing stimuli and accommodating multiple goals).

The second hypothesis predicted that game-based features should improve motivation and engagement during prolonged periods of training, which should have a corresponding increase in applied mastery (i.e., increased performance during later practice). The increase in self-explanation quality across Sessions 6 through 8 (Figure 10) lends support for this second hypothesis. Specifically, the increase in performance during these sessions corresponds with the improved affect and motivation ratings in the game-based condition (Figures 5–7).

Analyses on the self-explanation quality across time indicated that the game-based system resulted in a significant quadratic relation between training and performance, such that performance initially declined and subsequently increased. This curvilinear performance trajectory provides statistical support for both hypotheses and may also help to explain some of the mixed results found in the previous literature on educational games. Specifically, the time scale of measurement within a study may determine whether performance trends for game-based systems appear to be positive, negative, or neutral.

It is also worth noting that the minimal game features remaining in the ITS (see Table 1) were not enough to produce the same motivational improvements as the fully game-based version of training. This finding is potentially significant for two reasons. First, the fully game-based training likely would have produced even larger motivational differences if it had been compared with a more stripped-down version of an ITS (i.e., exaggerating the already significant effects). Second, just adding in a few game-like features to an ITS is not enough to produce the effects found in more coherent and contextually bound educational games. Our findings in this study demonstrate that the combined set of features and mechanisms integrated within our game-based system (feedback, incentives, control, task difficulty, and environment) effectively enhanced users' experience with the tutoring system and that most of these benefits tended to remain stable or even increase across time. The overall findings indicate that game-based inter-

action mechanisms can provide enjoyable, effective interactions that promote sustained motivation and mastery over time.

The current results further suggest that future research on educational games should incorporate multiple time scales of measurement to investigate the complex trajectories of both learning and motivation within these environments. As suggested from the current work, along with results from Jackson, Dempsey, et al. (2011, 2012), isolated measurements solely at pretest and posttest may provide an oversimplified snapshot of the potential benefits (or weaknesses) of game-based education. The current work is intended to inform researchers' future development and evaluation, as well as contribute to the need for empirical comparisons between game-based and nongame-based tutoring environments (O'Neil & Fisher, 2004; O'Neill et al., 2005).

The outcomes and concepts discussed here provide unique insight into various time-based effects within educational games. Repeated observations allow us to represent students' experiences throughout the interaction process and are further supported with more summative measures collected separately from training (i.e., posttest and delayed retention). The results from this study as well as our previous studies (Jackson, Dempsey, et al., 2011, 2012) support the assumption that students prefer working with game-based tutoring environments and that, over time, these systems can provide enjoyable training that produces learning outcomes comparable to more traditional ITSs. The current work provides substantial support for incorporating games into long-term tutoring environments and should help researchers and educators to better understand the potential benefits from these game-based components and systems.

References

- Adams, D. M., Mayer, R. E., MacNamara, A., Koenig, A., & Wainess, R. (2012). Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of Educational Psychology, 104*, 235–249. doi:10.1037/a0025595
- Alexander, P. A., Murphy, P. K., Woods, B. S., Duhon, K. E., & Parker, D. (1997). College instruction and concomitant changes in students' knowledge, interest, and strategy use: A study of domain learning. *Contemporary Educational Psychology, 22*, 125–146. doi:10.1006/ceps.1997.0927
- Anderson, J. R. (1982). Acquisition of a cognitive skill. *Psychological Review, 89*, 369–406. doi:10.1037/0033-295X.89.4.369
- Anderson, J. R., Corbett, A. T., Koedinger, K., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*, 167–207. doi:10.1207/s15327809jls0402_2
- Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Meheranian, H., Fisher, D., . . . Woolf, B. P. (2007). Repairing disengagement with non-invasive interventions. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Proceedings of the 2007 conference of artificial intelligence in education: Building technology-rich learning contexts that work* (pp. 195–202). Amsterdam, the Netherlands: IOS Press.
- Asgari, M., & Kaufman, D. (2004, September). *Intrinsic motivation and game design*. Paper presented at the 35th annual conference of the International Simulation and Gaming Association (ISAGA) and Conjoint Conference of SAGSAGA. Munich, Germany, September 6–10, 2004.
- Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In J. C. Lester, R. M. Vicari, and F. Paraguacu (Eds.), *Intelligent tutoring systems: 7th International conference, ITS 2004, Maceió, Brazil, August 30–September 3, 2004, Proceedings* [Lecture Notes in Computer Science 3220] (pp. 531–540). New York, NY: Springer.

- Baker, R. S. J. d., 'D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence and persistence of affect during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68, 223–241. doi:10.1016/j.ijhcs.2009.12.003
- Bandura, A. (2000). Self-efficacy: The foundation of agency. In W. Perig & A. Grob (Eds.), *Control of human behavior, mental processes, and consciousness: Essays in honor of the 60th birthday of August Flammer* (pp. 17–33). Mahwah, NJ: Erlbaum.
- Barab, S. A., Dodge, T., Ingram-Goble, A., Peppler, K., Pettyjohn, P., Volk, C., & Solomou, M. (2010). Pedagogical dramas and transformational play: Narratively rich games for learning. *Mind, Culture, and Activity*, 17, 235–264. doi:10.1080/10749030903437228
- Bell, C., & McNamara, D. S. (2007). Integrating iSTART into a high school curriculum. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society, Nashville, Tennessee, August 1–4, 2007* (pp. 809–814). Austin, TX: Cognitive Science Society.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13, 145–182. doi:10.1207/s15516709cog1302_1
- Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439–477.
- Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19, 237–248.
- Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence*, 16, 555–575. doi:10.1080/08839510290030390
- Corbett, A. T., & Anderson, J. R. (1990). The effect of feedback control on learning to program with the Lisp Tutor. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society, Cambridge, Massachusetts, July 25–28, 1990* (pp. 796–803). Hillsdale, NJ: Erlbaum.
- Corbett, A. T., & Anderson, J. R. (2001). Locus of feedback control in computer-based tutoring: Impact of learning rate, achievement, and attitudes. In M. Beaudouin-Lafon & R. J. K. Jacob (Eds.), *Proceedings of ACM CHI-2001 Conference on Human Factors in Computing Systems, Seattle, Washington, March 31–April 5, 2001* (pp. 245–252). New York, NY: ACM Press.
- Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88, 715–730. doi:10.1037/0022-0663.88.4.715
- Dempsey, K. B. (2011). *The effects of games on engagement and performance in intelligent tutoring systems* (Unpublished doctoral dissertation). Department of Psychology, University of Memphis, Memphis, Tennessee.
- D'Mello, S. K., Taylor, R., & Graesser, A. C. (2007). Monitoring affective trajectories during complex learning. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 203–208). Austin, TX: Cognitive Science Society.
- du Boulay, B. (2011). Towards a motivationally-intelligent pedagogy: How should an intelligent tutor respond to the unmotivated or the demotivated? In R. A. Calvo & S. 'D'Mello (Eds.), *New perspectives on affect and learning technologies* (pp. 41–52). New York, NY: Springer. doi:10.1007/978-1-4419-9625-1_4
- Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments*, 8, 111–127. doi:10.1076/1049-4820(200008)8:2;1-B;FT111
- Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, 33, 441–467. doi:10.1177/1046878102238607
- Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York, NY: Palgrave Macmillan. doi:10.1145/950566.950595
- Gee, J. P. (2005). *Why video games are good for your soul: Pleasure and learning*. Melbourne, Victoria, Australia: Common Ground Press.
- Graesser, A. C., Chipman, P., Leeming, F., & Biedenbach, S. (2009). Deep learning and emotion in serious games. In U. Ritterfeld, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 81–100). New York, NY: Routledge, Taylor, & Francis.
- Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40, 223–234. doi:10.3102/0013189X11413260
- Graesser, A. C., McNamara, D. S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point& Query, AutoTutor, and iSTART. *Educational Psychologist*, 40, 225–234. doi:10.1207/s15326985ep4004_4
- Gredler, M. E. (2004). Games and simulations and their relationships to learning. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 571–581). Mahwah, NJ: Erlbaum.
- Harris, D. (2008). A comparative study of the effect of collaborative problem-solving in a massively multiplayer online game (MMO) on individual achievement. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, 69(6-A), 2117.
- Jackson, G. T., Boonthum, C., & McNamara, D. S. (2010). The efficacy of iSTART extended practice: Low ability students catch up. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems, ITS 2010* [Lecture Notes in Computer Science 6094] (pp. 349–351). Berlin/Heidelberg, Germany: Springer.
- Jackson, G. T., Davis, N. L., Graesser, A. C., & McNamara, D. S. (2011). Students' enjoyment of a game-based tutoring system. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education: Proceedings of the 15th International Conference on Artificial Intelligence in Education, Auckland, New Zealand, June 28–July 2, 2011* [Lecture Notes on Artificial Intelligence 6738] (pp. 475–477). Berlin/Heidelberg, Germany: Springer.
- Jackson, G. T., Dempsey, K. B., Graesser, A. C., & McNamara, D. S. (2011). Short- and long-term benefits of enjoyment and learning within a serious game. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education: Proceedings of the 15th International Conference on Artificial Intelligence in Education, Auckland, New Zealand, June 28–July 2, 2011* [Lecture Notes on Artificial Intelligence 6738] (pp. 139–146). Berlin/Heidelberg, Germany: Springer.
- Jackson, G. T., Dempsey, K. B., & McNamara, D. S. (2012). Game-based practice in reading strategy tutoring system: Showdown in iSTART-ME. In H. Reinders (Ed.), *Computer games* (pp. 115–138). Bristol, England: Multilingual Matters.
- Jackson, G. T., Graesser, A. C., & McNamara, D. S. (2009). What students expect may have more impact than what they know or feel. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. C. Graesser (Eds.), *Artificial intelligence in education: Building learning systems that care. From knowledge representation to affective modeling* (pp. 73–80). Amsterdam, the Netherlands: IOS Press.
- Jackson, G. T., Guess, R. H., & McNamara, D. S. (2010). Assessing cognitively complex strategy use in an untrained domain. *Topics in Cognitive Science*, 2, 127–137. doi:10.1111/j.1756-8765.2009.01068.x
- Jackson, G. T., & McNamara, D. S. (2011). Motivational impacts of a game-based intelligent tutoring system. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference, Palm Beach, FL, May 18–20, 2011* (pp. 519–524). Menlo Park, CA: AAAI Press.

- Johnson, W. L., & Valente, A. (2008). Collaborative authoring of serious games for language and culture. In Elysebeth Leigh (Ed.), *Proceedings of SimTecT 2008*. Lindfield, New South Wales, Australia: Simulation Industry Association of Australia.
- Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 61–77). New York, NY: Cambridge University Press.
- Laird, J. E., & van Lent, M. (2000). Human-level AI's killer application: Interactive computer games. *AI Magazine*, 22, 15–26.
- Malone, T., & Lepper, M. (1987). Making learning fun: A taxonomy of intrinsic motivations of learning. In R. E. Snow & M. J. Farr (Eds.), *Aptitude, learning, and instruction: Vol. 3. Cognition and affective process analyses* (pp. 223–253). Hillsdale, NJ: Erlbaum.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38, 43–52. doi:10.1207/S15326985EP3801_6
- McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, 38, 1–30. doi:10.1207/s15326950dp3801_1
- McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch, (Eds.), *Handbook of latent semantic analysis* (pp. 227–241). Mahwah, NJ: Erlbaum
- McNamara, D. S., Jackson, G. T., & Graesser, A. C. (2010). Intelligent tutoring and games (ITaG). In Y. K. Baek (Ed.), *Gaming for classroom-based learning: Digital role-playing as a motivator of study* (pp. 44–65). Hershey, PA: IGI Global. doi:10.4018/978-1-61520-713-8.ch003
- McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments & Computers*, 36, 222–233. doi:10.3758/BF03195567
- McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research*, 34, 147–171. doi:10.2190/IRU5-HDTJ-A5C8-JVWE
- McNamara, D. S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I. B. (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 397–421). Mahwah, NJ: Erlbaum.
- Merrill, D. C., Reiser, B. J., Ranney, M., & Trafton, J. G. (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *Journal of the Learning Sciences*, 2, 277–305. doi:10.1207/s15327809jls0203_2
- Meyer, B. J. F., & Wijekumar, K. K. (2011). Individualizing a web-based structure strategy intervention for fifth graders' comprehension of non-fiction. *Journal of Educational Psychology*, 103, 140–168. doi:10.1037/a0021606
- Moreno, R., & Mayer, R. E. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology*, 97, 117–128. doi:10.1037/0022-0663.97.1.117
- O'Neil, H. F., & Fisher, Y. C. (2004). A technology to support leader development: Computer games. In D. V. Day, S. J. Zaccaro, & S. M. Halpin (Eds.), *Leader development for transforming organizations* (pp. 99–121). Mahwah, NJ: Erlbaum.
- O'Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. *Curriculum Journal*, 16, 455–474. doi:10.1080/09585170500384529
- Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38, 1–4. doi:10.1207/S15326985EP3801_1
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66, 543–578.
- Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & Education*, 52, 1–12. doi:10.1016/j.compedu.2008.06.004
- Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego, CA: Academic Press. doi:10.1016/B978-012109890-2/50043-3
- Pintrich, P. R., & Schrauben, B. (1992). Student's' motivational beliefs and their cognitive engagement in classroom academic tasks. In D. Schunk & J. Meece (Eds.), *Student perceptions in the classroom: Causes and consequences* (pp. 149–183). Hillsdale, NJ: Erlbaum.
- Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Education and Psychological Measurement*, 53, 801–813. doi:10.1177/0013164493053003024
- Rieber, L. P. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research and Development*, 44, 43–58. doi:10.1007/BF02300540
- Shute, V. (2008). Focus on formative feedback. *Review of Educational Research*, 78, 153–189. doi:10.3102/0034654307313795
- Shute, V., & Towle, B. (2003). Adaptive e-learning. *Educational Psychologist*, 38, 105–114. doi:10.1207/S15326985EP3802_5
- Steinkuehler, C. A. (2006). Why game (culture) studies now? *Games and Culture*, 1, 97–102. doi:10.1177/1555412005281911
- Van Eck, R. (2006). Building intelligent learning games. In D. Gibson, C. Aldrich, & M. Prensky (Eds.), *Games and simulations in online learning research & development frameworks* (pp. 271–307). Hershey, PA: IGI Global. doi:10.4018/978-1-59904-304-3.ch014
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wolters, C. (1998). Self-regulated learning and college students' regulation of motivation. *Journal of Educational Psychology*, 90, 224–235. doi:10.1037/0022-0663.90.2.224
- Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Benedict, L., . . . Yukhymenko, M. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research*, 82, 61–89. doi:10.3102/0034654312436980
- Zimmerman, B. J., & Kitsantas, A. (1997). Developmental phases in self-regulation: Shifting from process to outcome goals. *Journal of Educational Psychology*, 89, 29–36. doi:10.1037/0022-0663.89.1.29
- Zimmerman, B. J., & Schunk, D. H. (2001). Reflections on theories of self-regulated learning and academic achievement. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp.289–307). Mahwah, NJ: Erlbaum.

Received December 15, 2011

Revision received January 14, 2013

Accepted February 11, 2013 ■

The Impact of Individual, Competitive, and Collaborative Mathematics Game Play on Learning, Performance, and Motivation

Jan L. Plass
New York University

Paul A. O’Keefe
New York University and The City University of New York

Bruce D. Homer and Jennifer Case
The City University of New York

Elizabeth O. Hayward, Murphy Stein, and
Ken Perlin
New York University

The present research examined how mode of play in an educational mathematics video game impacts learning, performance, and motivation. The game was designed for the practice and automation of arithmetic skills to increase fluency and was adapted to allow for individual, competitive, or collaborative game play. Participants ($N = 58$) from urban middle schools were randomly assigned to each experimental condition. Results suggested that, in comparison to individual play, competition increased in-game learning, whereas collaboration decreased performance during the experimental play session. Although out-of-game math fluency improved overall, it did not vary by condition. Furthermore, competition and collaboration elicited greater situational interest and enjoyment and invoked a stronger mastery goal orientation. Additionally, collaboration resulted in stronger intentions to play the game again and to recommend it to others. Results are discussed in terms of the potential for mathematics learning games and technology to increase student learning and motivation and to demonstrate how different modes of engagement can inform the instructional design of such games.

Keywords: achievement goal orientations, games, learning

The past decade has seen an intensifying interest in the use of digital games in pursuit of educational goals. Entertainment games, whether they run on a computer, game console, mobile device, or touch pad, are highly engaging and motivating, and educators have suggested taking advantage of these qualities of

games to facilitate learning (Gee, 2007; Kafai, 1995; Squire, 2003). Proponents of digital game-based learning have argued that well-designed games embody educational and learning theory and are in line with some of the “best practices” of education (e.g., Barab, Ingram-Goble, & Warren, 2008; Collins & Halverson, 2009; Gee, 2003; Mayo, 2007; Shaffer, 2008; Squire, 2008).

The validation of claims that good games are good for learning first leads us to consider the question: What is a good game? There are many aspects of the design of a digital game that can impact the game’s educational effectiveness. For example, game designers make decisions regarding the game’s core mechanic (Plass et al., in press; Salen & Zimmerman, 2003), the representation of the game content (Plass et al., 2009), the emotional design of the game (Um, Plass, Hayward, & Homer, 2012), the game’s incentive system, and social aspects of play (Salen & Zimmerman, 2003). Research on the design of good games for learning therefore examines the effects of key features of games on students’ learning experiences and outcomes (Plass, Homer, & Hayward, 2009). The goal of this line of research is to investigate whether effects of social, cognitive, and affective factors related to learning found in research on other learning environments can be extended to the design of games for learning and used to develop theory-based, empirically validated design patterns for such games. Design patterns, originally proposed in the context of architecture (Alexander, Ishikawa, & Silverstein, 1977), represent general solutions to commonly occurring problems that educational game designers can use to guide the design of specific aspects of their games.

In the present study, we examined one of these design patterns—the context of playing a game—to increase arithmetic flu-

This article was published Online First September 9, 2013.

Jan L. Plass, Games For Learning Institute (G4LI) and Consortium for Research and Evaluation of Advanced Technologies in Education (CREATE), New York University; Paul A. O’Keefe, G4LI and CREATE, New York University, and The Graduate Center, The City University of New York; Bruce D. Homer and Jennifer Case, G4LI and Program in Educational Psychology, The Graduate Center, The City University of New York; Elizabeth O. Hayward, Murphy Stein, and Ken Perlin, G4LI and Department of Computer Science, Courant Institute of Mathematical Sciences, New York University.

Paul A. O’Keefe is now at the Department of Psychology, Stanford University.

The work reported in this article was funded in part by a grant from Microsoft Research to the G4LI. The content and opinions herein are the authors’ and may not reflect the views of Microsoft Research, nor does mention of trade names, products, or organizations imply endorsement. Jan L. Plass and Paul A. O’Keefe contributed equally to this work. We thank the G4LI research assistants who assisted in the data collection for this research.

Correspondence concerning this article should be addressed to Jan L. Plass, New York University, CREATE, 196 Mercer Street Suite 800, New York, NY 10012, or Paul A. O’Keefe, Department of Psychology, Jordan Hall, Building 420, 450 Serra Mall, Stanford University, Stanford, CA 94305. E-mail: jan.plass@nyu.edu or paul.okeefe@stanford.edu

ency. Middle-school students were randomly assigned to play an arithmetic game, *FactorReactor*, developed by the *Games for Learning Institute* for the purpose of this research. They played either on their own (*individual*), against another student (*competitive*), or together with another student (*collaborative*). Learning, performance, achievement goal orientations, interest, enjoyment, and future game intentions were examined as a function of mode of play.

Theoretical Background

In the present research, we were interested in how three modes of play (individual, competitive, and collaborative) affect learning, game performance, and motivation. The conceptual framework for this research consists of the educational context of learning, achievement goal theory, and interest for which we review related research in this section.

Educational Contexts

It has long been established that social context generally, and peer interaction specifically, impact the learning process and that knowledge construction is a social, collaborative process (Light & Littleton, 1999; Piaget, 1932; Salomon, 1993; Scardamalia & Bereiter, 1991; Vygotsky, 1978). Research on the social context of learning has found that peer involvement in learning can affect both academic achievement as well as learner attitudes in a variety of contexts. Early work on cooperative learning in the classroom context suggests that peer collaboration may have positive effects on academic achievement across a variety of content areas (Berg, 1994; Dillenbourg, 1999; Slavin, 1980, 1983; Slavin, Leavey, & Madden, 1984). Cooperative learning has also been found to increase positive attitudes toward school generally and mathematics as a subject area (Slavin, 1980; Slavin et al., 1984). Research on competition suggests that learning and performance are better in competitive compared with individual settings (Ames, 1984) and that competitive features result in the development of analytic skills (Fu, Wu, & Ho, 2009), but not always in increased learning outcomes (Ke & Grabowski, 2007).

Collaboration

Group collaboration can take a variety of forms and has been investigated in a broad range of contexts, including classroom-based learning (Berg, 1994), computer-based learning (R. T. Johnson, Johnson, & Stanne, 1986; Mevarech, Stern, & Levita, 1987; Scardamalia & Bereiter, 1991), and web-based and e-learning (Hron & Friedrich, 2003). What these collaborations have in common is that two or more learners interact in a synchronous form to negotiate shared meaning and jointly and continuously solve problems (Dillenbourg, 1999).

The recent surge in interest in digital games as tools for learning offers up a new forum for investigating learning as a social activity. Initial research has provided thick descriptions and case studies of such collaborative activities in learning with games and related activities (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005; Squire, 2005; Steinkuehler, 2006). In comparison to individual study, group collaboration appears to be well suited for problem solving because collaboration encourages students to ex-

plain their thinking, verbalize it, and engage in joint elaboration on their decision making (Mullins, Rummel, & Spada, 2011). In addition, Kirschner, Paas, Kirschner, and Janssen (2011) showed that students working in small groups were better able to handle the cognitive load demands of problems with complex information, and thus learned more efficiently, than students solving problems in individual work.

The effects of collaboration only accrue, however, when certain conditions are met. In fact, a meta-analysis by Lou et al. (1996) found that collaboration did not have an effect in about one fourth of the studies, and in some cases even had a negative impact. Some of the conditions for the effectiveness of group collaboration are so fundamental that many consider them part of the definition of collaborative learning: Group members must have a shared group goal that they deem important, and the success of the activity must depend on all members of the group; that is, each member must be individually accountable (Slavin, 1988).

In addition to these fundamental conditions, additional ways to support group collaboration have been explored. Berg (1994), for example, used collaboration scripts to facilitate group collaboration, and Hron, Hesse, Cress, and Giovis (2000) showed that structuring the dialogue in group collaboration enhanced learners' orientation to the subject matter and reduced off-task conversation, though it did not increase knowledge gains. Other ways to assure the success of collaborative learning includes providing students with visualization tools (Fischer, Bruhn, Gräsel, & Mandl, 2002), managing the cognitive load they experience (van Bruggen, Kirschner, & Jochems, 2002), and providing adaptive support from intelligent tutors (Diziol, Walker, Rummel, & Koedinger, 2010) and from interactive dialogue agents (Chaudhuri et al., 2008).

The beneficial performance effects of collaboration only appear to be present for tasks involving conceptual knowledge, but not for procedural skill fluency (Mullins et al., 2011). In their research, Mullins et al. (2011) found that collaboration improved learning for both conceptual and procedural (skill fluency) material but that students in the procedural skill task engaged in ineffective learning behaviors. This is supported by other studies of group collaboration on learning involving conceptual knowledge that found that students provide explanations to one another (Diziol, Rummel, Spada, & McLaren, 2007) and engage in joint elaboration and co-construction of knowledge (Berg, 1994). The same kind of elaboration was not found in procedural skills acquisition.

In the present study, we were interested in investigating collaboration on a game-based task of arithmetic fluency development. Even though research so far has not shown clear benefits of collaboration for skills automation, other research suggests that conceptual knowledge and skills acquisition are linked, and the development of one can benefit the other (Rittle-Johnson & Alibali, 1999; Rittle-Johnson, Siegler, & Alibali, 2001).

Arithmetic skills development begins in early childhood and continues throughout formal and informal schooling with the goal of becoming automated, but even adults often still use strategies to solve basic problems of addition, subtraction, multiplication, and division rather than retrieving basic arithmetic facts from long-term memory (Tronsky, 2005). The adaptive strategy choice model developed by Siegler and colleagues (Lemaire & Siegler, 1995; Shrager & Siegler, 1998) describes the development of strategy use along four dimensions as arithmetic experience increases. These dimensions include (a) which strategies are available to the

learner, (b) when a particular strategy is used, (c) how that strategy is executed, and (e) the decisions governing which strategy is chosen. As learners encounter arithmetic problems, they select and carry out a strategy to solve the problem and, in the process, accumulate data on the effectiveness of their strategy on multiple levels (Shrager & Siegler, 1998). Research has shown that, over time, automation (i.e., retrieval of the correct answer from memory) becomes the dominant strategy because it yields highest accuracy rates and shortest response times (Tronsky, 2005).

FactorReactor was designed to support this skill automation in middle-school-age children by providing arithmetic problems that increase in difficulty from one level to the next. Small-group collaboration was found to be beneficial in the classroom even for the development of arithmetic skills (Yackel, Cobb, & Wood, 1991), and we were interested in whether a collaborative mode in the game would result in higher performance compared with an individual play mode.

Competition

A common element of video games is a competitive mode in which players compete with one another. In some cases, this competition means that two or more players compete for the same goal, such as in the table tennis game in *Wii Sports Resort*. In other cases, both players play the same game individually but are aware of each other's progress and score, such as in the bowling game in *Wii Sports Resort*.

Many studies investigating the effect of competitive forms of learning compare various social modes with an individual mode. A meta-analysis of 122 studies, comparing the effects of individual, competitive, and collaborative goal structures on achievement, found benefits for collaborative compared with competitive or individual goal structures (D. W. Johnson, Johnson, Maruyama, Nelson, & Skon, 1981). In a related study, R. T. Johnson, Johnson, and Stanne (1986) compared the effect of computer-assisted cooperative, competitive, and individual learning on performance and attitudes. Eighth graders were randomly assigned to work in either a small group, in the cooperative and competitive conditions, or individually to learn about fundamentals of map reading and navigation. Students in the cooperative condition were found to show the highest performance on daily worksheets. However, both the cooperative and competitive groups had higher levels of interest in computers at the close of the study, as compared with those who worked individually. More recently, Fu et al. (2009) investigated the knowledge creation process in a web-based learning environment concerning computer software. The authors predicted that the social presence of peers, in the form of a partner, would increase performance as well as enjoyment motivation. Four conditions were compared in which the collaborative (presence vs. absence of a partner) and competitive (presence vs. absence of financial reward and grade feedback) features of group learning of undergraduate students were systematically varied. Results indicate that both the collaborative and competitive features increased enjoyment in learning. When competitive features were present, students demonstrated higher analytic skills, or the separation of concepts into component parts as a means to understand organizational structure, as defined by Bloom's (1956) taxonomy. The collaborative feature encouraged higher synthetic skills, or the building of structure from information, and therefore was indica-

tive of higher level learning. The authors concluded that both collaborative and competitive elements worked to bolster performance in a web-based environment.

Strommen (1993) compared cooperative and competitive contexts in learning from a computer-based natural science game among fourth graders. Students in the collaborative condition were found to be more successful in their game performance and used more game play strategies as compared with those in the competitive condition. Ke and Grabowski (2007) used a math computer game addressing measurement, whole numbers, equations, and graphing to examine the impact of cooperative game play, individual play, and competitive game play in fifth graders. After eight 40-min game play sessions, there was no difference between the cooperative and competitive conditions in achievement, as measured by multiple-choice arithmetic test. However, students in the cooperative condition demonstrated more positive math attitudes at the close of the study as compared with those in the competitive condition, further suggesting that the presence of peers when learning impacts attitudes toward academic content.

Because competition is a common element of games, and because some research suggests that performance is better in competitive compared with individual settings (Ames, 1984), we were interested in how learning and performance in the competitive play version of *FactorReactor* compared with individual play.

Achievement Goal Orientations

The structure of learning environments and the tasks used to engage learners can elicit particular achievement goals that can either facilitate or hinder learning (Ames, 1992; Meece, Anderman, & Anderman, 2006). Similarly, modes of play may influence the adoption of particular goal orientations. Achievement goal theory posits two major types of goal orientations people endorse in achievement situations: *mastery* and *performance* (Ames & Archer, 1988; Dweck & Leggett, 1988; Elliot, 2005). A mastery goal orientation focuses on learning and the development of abilities, and success is defined in terms of personal improvement. In contrast, performance goal orientations focus on demonstrating or validating abilities, and success is defined in terms of performing well compared with others (Elliot, 1999, 2005). It is distinct from competition, however, in that outperforming others is a means of demonstrating or validating abilities rather than being the goal in and of itself. Performance goals can further be subdivided into approach and avoidance dimensions (Cury, Elliot, Da Fonseca, & Moller, 2006; Elliot, 2005; Elliot & McGregor, 2001). A performance-approach goal orientation focuses on performing well compared with others, whereas a performance-avoidance goal orientation is concerned with evading the appearance of incompetence and performing poorly relative to others. This approach-avoidance distinction has also been made with regard to mastery goals (Elliot, 1999; Elliot & McGregor, 2001); however, there is less empirical support for it (Maehr & Zusho, 2009). Therefore, we used the trichotomous model in the present research, assessing mastery-approach (which we refer to as *mastery*), performance-approach, and performance-avoidance goals among learners.

In general, research has found that mastery goal orientations result in highly adaptive patterns of motivation and learning (Midgley, Kaplan, & Middleton, 2001). For example, they are associated with high levels of effort and persistence (Grant &

Dweck, 2003), particularly on difficult tasks (Elliott & Dweck, 1988; Stipek & Kowalski, 1989), increased task involvement (Harackiewicz, Barron, Tauer, Carter, & Elliot, 2000), and increased self-efficacy (Meece, Blumenfeld, & Hoyle, 1988; Midgley et al., 1998). Moreover, mastery goal orientations are associated with enhanced learning strategies that lead to better understanding of concepts and recall (Ames & Archer, 1988; Elliot & McGregor, 2001; Grant & Dweck, 2003). Although performance-approach goals can also have adaptive outcomes, such as high academic achievement (Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002), these benefits can be accompanied by test anxiety (Linnenbrink, 2005; Skaalvik, 1997), cheating (Tas & Tekkaya, 2010), and the avoidance of help seeking (Karabenick, 2004). In contrast, performance-avoidance goals are consistently found to result in maladaptive motivational outcomes (Elliot & Mapes, 2005; Harackiewicz et al., 2002; Midgley et al., 2001). They are associated with lower achievement, intrinsic motivation, academic self-efficacy, and engagement (e.g., Church, Elliot, & Gable, 2001; Elliot & McGregor, 1999; Middleton & Midgley, 1997; Pekrun, Elliot, & Maier, 2009; Skaalvik, 1997).

Taken together, mastery goal orientations provide the most adaptive framework from which to pursue educational goals, and contexts structured to invoke these goals have the potential to benefit student motivation in the long run. For example, O'Keefe, Ben-Eliyahu, and Linnenbrink-Garcia (2013) found that a mastery-structured learning environment not only attenuated students' performance-approach and -avoidance goal orientations but also augmented mastery goal orientations. Furthermore, the observed increases in mastery goal orientations were sustained 6 months after students had returned to more traditional, performance-oriented learning environments.

In the present research, we examined how playing an educational game by oneself, in competition with another, or collaboratively results in the adoption of various achievement goal orientations. Given their influence on the adaptiveness of motivational and learning patterns, in the present study we intended to shed light on how the design and implementation of educational games can result in optimal motivational outcomes. A study by Ames (1984) found that working individually on a set of puzzles led children to attribute their level of performance to the effort they had expended, whereas those working competitively attributed their performance to their level of ability. Given that these attributional patterns map onto mastery and performance goal orientations, respectively (Dweck, 1986; Dweck & Leggett, 1988), we might expect that performance goal orientations would be adopted more strongly in the competitive condition relative to the individual play condition. The context of a game, however, may change the meaning of competition. Although games can heighten concerns about performance, they do not necessarily heighten concerns about the demonstration or validation of normative ability. Instead, educational games, such as the one employed in the present research, are designed to produce incremental personal success, which is in line with a mastery goal orientation. Therefore, we expected that playing competitively would increase mastery goal orientations as compared to individual play and that performance goal orientations would not be affected by the competitive game context.

We expected the collaborative condition to have a similar effect on players' mastery goal orientations. A study by Ames and Felker (1979) examining children's attributions regarding the achieve-

ment outcomes of another student found that ability attributions were stronger for those who worked individually and competitively than collaboratively. Similarly, effort attributions were stronger for individual and competitive work than successful (as compared with unsuccessful) collaborations. These results would suggest an ambiguous prediction regarding the adoption of achievement goal orientations in the types of contexts that are traditionally examined. However, the context of an educational game is different than the nongame contexts that are typically studied, largely because it provides a framework for incremental personal improvement. Accordingly, we expected that collaborative play would invoke stronger mastery goal orientations compared to individual play, and that performance goal orientations would remain unaffected.

Interest

Mode of play should similarly have an impact on players' interest in the game. First, it is useful to distinguish two general types of interest. *Individual interest* refers to an intrinsic desire and tendency to engage in particular ideas, content, and activities over time. For example, someone with an individual interest in sports may watch games on television, read up on player stats, or play in a competitive athletic league, and engage in these activities on a relatively regular basis. *Situational interest*, in contrast, refers to the attentional and affective reactions elicited by the environment (e.g., Hidi & Renninger, 2006; Linnenbrink-Garcia et al., 2010). For instance, a physics instructor explaining how rockets work may not elicit much situational interest in his or her students using traditional lecture methods; however, he or she would likely elicit high situational interest by having students build and launch their own rockets. Although situational interest involves elements that include feelings of excitement and fascination, it is distinct from other constructs, such as enjoyment, in that it also includes elements relating to the personal value of the interest object or involvement in the activity.

Situational interest is of particular importance in education because it is essential to the development of individual interest. According to Hidi and Renninger's (2006) four phase model, once situational interest is triggered, it can be maintained when personal relevance or involvement is established. Individual interest begins to emerge when the individual develops a relatively persistent predisposition to reengage in particular ideas, content, or activities. Finally, well-developed individual interest emerges once contextual supports are no longer necessary, such that the interest is generally, but not exclusively, self-generated. In the present study, we were interested in how the modes of play, particularly competitive and collaborative, influence situational interest, as it may suggest how games for learning can be designed and implemented to effectively elicit situational interest, and ultimately develop into individual interest in academic topics.

Although one of the defining characteristics of games is to elicit situational interest (Salen & Zimmerman, 2003), the extent to which individual, competitive, and collaborative modes of play contribute to its invocation has not yet been examined experimentally. We expected that competitive and collaborative play modes would elicit greater situational interest than playing alone due to the social aspect of playing against or with a partner. These social contexts should enhance the excitement of game play, as well as

personal involvement. Additionally, we expected that other indicators of interest and motivation would reflect this prediction, such that the competitive and collaborative conditions should lead to greater enjoyment of the game, as well as a greater likelihood of future game reengagement and recommending the game to others.

The Present Study

In the present study, we aimed to investigate how three modes of play (individual, competitive, and collaborative) affect learning, game performance, and motivation. As discussed above, social educational contexts, such as competition and collaboration, have been shown to affect learning in a variety of settings, such as classrooms and web-based environments, for different age groups, and for different levels of learning objectives. It is of great interest to game designers and motivation theorists alike whether similar effects can be found for digital games designed for educational purposes. We, therefore, investigated how competitive and collaborative modes of play compared with individual play in impacting learning, performance, achievement goal orientations, situational interest, enjoyment, and intentions to reengage in the game and recommend it to others. Our focus on these outcomes reflects the important intentions of using games for educational purposes, such that they have the potential to improve performance and increase engagement in educational activities. For the present research, we used *FactorReactor*, a game designed to practice and automate arithmetic skills to increase arithmetic fluency in middle-school-age students.

Method

Participants. Participants were 58 sixth-, seventh-, and eighth-grade students (58.6% female) from seven urban public schools in a major northeastern city. All students were taking part in a technology-themed afterschool program led by a teacher at their school. Membership in each of the programs was small and voluntary. In partnership with these programs, researchers made weekly visits to each school during the academic year to introduce students to educational technologies and games. In one of the sessions, students participated in the present study. The mean age of the students was 11.02 years ($SD = 3.61$). Missing data were handled through listwise deletion.

Procedure. Before students arrived to the classroom in which the study was run, tables were arranged so that computer stations could be set up sufficiently far apart from one another. When students arrived, their assent and parental consent was collected, and were then seated at a computer station. They first watched an instructional video on their computers that provided an overview of the rules and goals of the game, *FactorReactor*, as well as how to use the Xbox game controller. Computer monitors were either 13 or 15 in. (33 or 38.1 cm). All participants then played a practice round of the game individually for 5 min. During this time, they were provided with a controller schematic sheet to assist in learning the operation of the controller. At the end of the practice session, an experimenter was available to the students to clarify any issues regarding the game, and the controller schematic sheet was taken away. Next, all participants played the game independently for 3 min, which constituted the pretest of game performance.

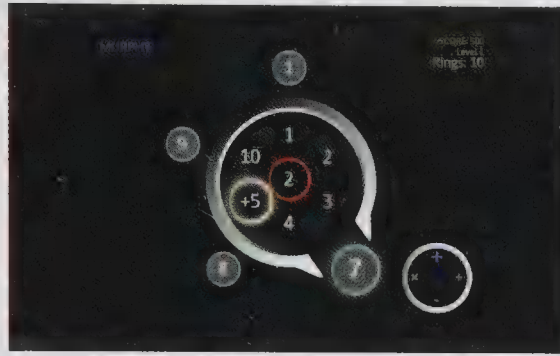
Students were then randomly assigned to one of three modes of play: *individual* ($n = 16$), *competitive* ($n = 20$), and *collaborative* ($n = 22$). This also meant that partners in the competitive and collaborative conditions were random. Participants in the individual condition were situated in front of a laptop computer with a single controller, whereas the competitive and collaborative conditions joined with a partner in front of a laptop computer with two controllers. Before beginning, an experimenter provided the context for the experimental game play and specific instructions to the students. Those in the individual condition were told that they would be playing the same version of the game as before and were given the following instructions: "When playing the game, get the best score you can." Those in the competitive condition were told that they would be playing a version of the game that allowed two players to compete against each other and were given the following instructions: "When playing the game, compete against each other for the better score." Those in the collaborative condition were told that they would be playing a version of the game that allowed two players to play together and were given the following instructions: "When playing the game, work together to get the best score." Instructions to learners regarding how to collaborate were kept relatively short for three reasons: First, middle-school-age students are used to playing games without receiving elaborate instructions and would likely have skipped any instructions provided to them. Second, models of mathematics learning describe students as active learners who spontaneously create their own strategies to solve a problem (Cobb, Wood, & Yackel, 1991), and we did not want to stifle this invention of strategies by prescribing the process of collaboration. Finally, critical reviews of studies involving various forms of scaffolding have argued that performance differences between the individual and collaborative group found in such studies could have been attributable to the fact that the scaffolding (elaboration scripts, dialogue scaffolding, visualizations) was only given to the collaborative group (Mullins, Rummel, & Spada, 2010).

Participants were given 15 min to play, at which point the game automatically stopped. Figure 1 shows screen shots of the game in the three play modes. Participants then played another 3-min individual play session as a posttest of game performance. At the end of game play, participants were independently administered surveys assessing game-relevant achievement goal orientations, situational interest in the game, game enjoyment, future intentions regarding the game, and their degree of experience with video game controllers. Finally, they completed another individual 3-min play session.

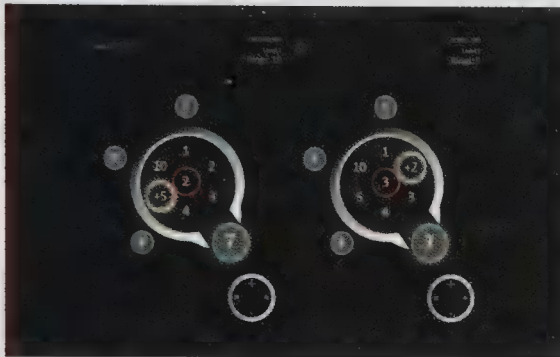
Materials

FactorReactor. *FactorReactor* is a game designed to practice and automate arithmetic skills, and was adapted from the original version to investigate cognitive and motivational outcomes related to mode of play. The game runs on a PC and is played with an Xbox controller connected to the PC via USB cable. Figure 1 shows screen shots of the game for each mode of play. Arithmetic fluency was chosen because it was identified by many teachers in our collaborating middle schools as a key skill on which other skills from the common core standards in Grades 6–8 build, but which is not sufficiently developed in many middle-school students.

A) Single Player



B) Competitive



C) Collaborative

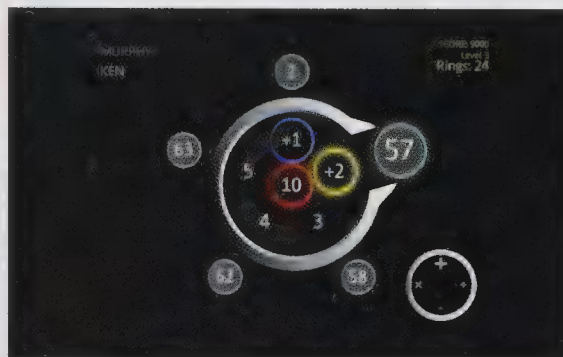


Figure 1. Three modes of play in *FactorReactor*: A: individual play. B: competitive play. C: collaborative play. *FactorReactor* by Murphy Stein and Games for Learning Institute is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 unported license.

FactorReactor possesses the key defining elements of a game: It has a clear goal and clear rules of play, has an engaging game mechanic that allows for a high degree of player choice, provides feedback and incentives, and has a fail state (Salen & Zimmerman, 2003). The object of the game was to transform the center number into one of the surrounding goal numbers by adding, subtracting, multiplying, or dividing it by one of the numbers from the inner ring. This was conducted by selecting one of the operators (+, −, ×, ÷) and one of the inner numbers and then hitting the *fire* button. For example, if the goal number was 7 and the center number was 2, as it is in Figure 1A, the player might select 5 from the inner ring, and then choose the “+” operator. By pressing the *fire* button, 5 would be added to the center number, transforming it to 7. A subsequent press of the *fire* button then solved the problem and automatically advanced to the next goal number. When transformations were done correctly, the center ring turned from red to green. If incorrect or impossible transformations were attempted (e.g., dividing the center number so that it does not result in a whole number, such as $2 \div 5$), the center number would temporarily glow and jiggle. Players had full control over which goal number they worked on at any given time, however, affording considerable flexibility in solving each problem.

Each time the center number was correctly transformed, the player earned a token, called a “ring,” and each player began the game with 10 rings. The number of rings rewarded for a correct transformation was equal to 2 times the minimum possible number of transformations for the relevant solution. Players could be awarded between two and eight rings, depending on the problem

that was solved (i.e., problems required, at a minimum, between one and four steps to be solved). Rings are used up with each operation, such that when a player hits the *fire* button, a ring is used; therefore, if a player attempts to transform the center number using multiple operations, multiple rings are used. In this way, the game disincentivized players from reaching their goal by guessing or repeating the same simple operation again and again (e.g., repeatedly subtracting a small number) and encouraged them to use more complex operations to solve problems in fewer moves. Scores were also calculated, which were highest for those who solved each problem using the least amount of rings. The level ended when all goal numbers were computed properly and at least one ring remained.

Levels increased in difficulty, such that the operations needed to reach the goal number became more complex. For example, a player may not be able to simply subtract an inner-ring number to successfully transform the center number as in easier levels. They may instead need to divide by one inner-ring number and then add another inner-ring number, or perhaps a more complex series of transformations. When a player ran out of rings, they received a “Game Over” message and were required to start the current level from the beginning. These messages were therefore an indicator of the use of inefficient strategies used by the players to solve the arithmetic problems presented by the game.

The game screen for the individual and collaborative play conditions were nearly identical (see Figure 1A and 1C). They had one game interface, which included one center number, five inner-ring numbers, and five goal numbers. The only difference between the

two was that, in the collaborative condition, both players had simultaneous and independent control over the game operations. That is, each player could select operators, inner-ring numbers, goal numbers, and also hit the *fire* button. Furthermore, player names were displayed in the upper-left portion of the screen, and in the upper-right portion of the screen were indicators of game performance, which included their current score, level, and number of rings. In the competitive condition, players had their own game interfaces, which were placed side-by-side (see Figure 1B). Each interface was identical to the individual play condition; however, indicators of each player's game performance were present and visible to both players. Furthermore, both players could work at their own pace, independently advancing through the levels.

Measures

Within-game learning and performance measures. Two indicators of game performance were used. Within-game learning was assessed with the total number of problems solved during the posttest individual game play period. During this game period, players were presented with problems on a similar level of difficulty as during the pretest and experimental sessions. The number of problems they solved, and the challenge level they reached, depended on how fast they progressed in the game. Increased performance during the posttest should suggest that arithmetic learning had occurred during the experimental session. The other indicator was the total number of "Game Over" messages players received during the experimental game play. When a player ran out of rings, he or she received a message stating, "You ran out of rings. The FactorReactor was destroyed," and then players restarted the level, which contained the same problems. Players ran out of rings either because they failed to solve any problems correctly, thereby failing to earn rings, or because they were not efficient enough in solving the problems. Therefore, the number of times a player received this "Game Over" message was also considered indicative of game performance. Pretest performance for each indicator was also collected and used as covariates in the analyses.

Achievement goal orientations. Participants were given the Achievement Goal Orientation subscale from the Patterns of Adaptive Learning Scales (Midgley et al., 2000). The language in the scale was simplified to ensure comprehension in our middle-school sample and was adapted to be relevant for game play. The 14-item survey asked students to indicate their level of agreement using a 7-point scale (1 = *Very much disagree*, 4 = *Neither agree nor disagree*, 7 = *Very much agree*) in response to items such as "One of my goals was to learn as much as I could about the game" (mastery; $\alpha = .87$), "One of my goals was to show others that the game was easy for me" (performance-approach; $\alpha = .84$), and "It was important to me that my performance on the game didn't make me look stupid" (performance-avoidance; $\alpha = .70$).

Situational interest. Situational interest was measured using an adaptation of the Situational Interest Survey (Linnenbrink-Garcia et al., 2010). The language of the survey was simplified to ensure comprehension in our middle-school sample and was adapted to be relevant for game play. The survey assessed several aspects of situational interest, including affective responses to the game (e.g., excitement, fascination) and its personal importance. Participants used a 7-point scale anchored at 1 (*Very much dis-*

agree), 4 (*Neither agree nor disagree*), and 7 (*Very much agree*) to indicate their level of agreement with 12 statements, such as "The game was exciting," "I learned valuable things from the game," and "What I learned from the game is fascinating to me" ($\alpha = .92$).

Game enjoyment. Overall enjoyment of the game was assessed with two questions asking participants to rate the extent to which they had fun playing the game and how much they liked the game, on a 5-point scale anchored at 1 (*Not at all*) and 5 (*A lot*) ($\alpha = .80$).

Future game intentions. Two items assessed participants' future intentions regarding *FactorReactor*. The first assessed intentions to reengage in the game, asking "Would you play this game again in the future?" The other assessed their intention to recommend the game to someone else, asking "Would you recommend it to your friends/teachers?" Both items were assessed on a 5-point scale ranging from 1 (*Not at all*) to 5 (*Definitely*).

Prior experience with video game controllers. Participants were asked to indicate their level of experience with video game controllers like the ones used in the study, rated on a 5-point scale anchored at 1 (*None*) and 5 (*A lot*). This variable was used as a covariate on the game performance analyses (see Table 1).

Out-of-game learning measure. Participants were given a pre- and posttest of math fluency as an out-of-game assessment of arithmetic learning. The measure included 160 simple arithmetic problems for which participants were given 3 min to complete as many problems as possible. This measure of math fluency was adapted from the *Woodcock-Johnson III Math Fluency subtest* (McGrew & Woodcock, 2001), modified by randomizing the presentation of problems and by including simple division problems as well as addition, subtraction, and multiplication problems. The posttest of math fluency was identical to the pretest, though the problems were presented in a different, randomized order to diminish practice effects.

Results

The data were analyzed using hierarchical linear models (HLMs). In these models, individuals were nested within pairs for the sole purpose of accounting for the correlated variance between individuals playing in dyads, which was the case for two of three of the experimental conditions (i.e., competitive and collaborative play). The main intention of our analyses, however, was to draw conclusions at the level of the individual, not the pairs level, so our report chiefly focuses on individual-level effects.

Across all analyses, mode of play was dummy coded with competitive play and collaborative play entered into the models, and individual play as the reference group. All analyses were run using HLM Version 7 (Raudenbush, Bryk, & Congdon, 2011). No gender or grade-level differences were found for the dependent variables; therefore, gender and grade level are not considered in further analyses.

Game Performance

Two indicators of game performance were analyzed. In our first analysis, we examined the effect of mode of play (individual vs. competitive vs. collaborative) on the number of problems solved in the posttest of game play. We ran an HLM with number of problems solved as the dependent variable, the two mode of play

Table 1
Descriptive Statistics and Correlations for Dependent Variables and Covariates

Variable	Possible range	N	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Game pretest score	0–	52	5.54	3.93	—													
2. Game posttest score	0–	52	10.06	5.07	.47***	—												
3. Pretest Game Over messages	0–	43	.84	.81	.24	.19	—											
4. Experimental session Game Over messages	0–	49	2.67	3.02	-.08	-.06	.38*	—										
5. Pretest math fluency	0–160	58	67.22	26.34	.26*	.41***	-.25	-.45***	—									
6. Posttest math fluency	0–160	57	70.42	26.67	.19	.30*	-.31*	-.43**	.92***	—								
7. Situational interest	1–5	58	5.46	1.06	-.04	.07	.10	-.07	.05	.00	—							
8. Game enjoyment	1–5	57	4.04	.80	-.03	.10	.17	.11	-.01	-.04	.80***	—						
9. Reengagement intention	1–5	55	3.95	.95	-.08	-.04	.18	.06	-.12	-.13	.68***	.69***	—					
10. Recommendation intention	1–5	55	3.93	1.15	-.10	-.09	.15	.14	-.28*	-.29*	.69***	.63***	.69***	—				
11. Mastery goal orientation	1–7	55	5.62	1.20	-.07	.08	.14	-.07	-.01	.02	.77***	.59***	.52***	.52***	—			
12. Performance-approach goal orientation	1–7	55	4.20	1.58	.06	.11	.07	.12	.03	.02	.28*	.21	.29*	.30*	.32*	—		
13. Performance-avoidance goal orientation	1–7	55	4.12	1.55	.11	.18	-.06	.14	.06	-.03	.20	.17	.11	.02	.30*	.71***	—	
14. Prior controller experience	1–5	53	4.21	1.03	.13	-.06	.27	.31*	-.28*	-.29*	.06	.12	.04	.01	-.10	.06	.00	—

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

condition dummy variables, and two covariates: The number of pretest play problems solved served as a baseline of game ability, and the degree of players' experience with video game controllers served as a baseline for previous experience with controller-based video games. The equations took the following forms:

$$\text{Level 1: } Y_{ij}(\text{posttest score}) = \beta_{0j} + \beta_{1j}(\text{competition})$$

$$+ \beta_{2j}(\text{collaboration}) + \beta_{3j}(\text{pretest score})$$

$$+ \beta_{4j}(\text{controller experience}) + r_{ij}.$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \beta_{1j} = \gamma_{10}, \beta_{2j} = \gamma_{20}, \beta_{3j} = \gamma_{30}, \beta_{4j} = \gamma_{40}.$$

Results showed a fixed effect for the competitive condition ($p = .02$), such that players in that condition performed significantly better than those in the individual condition (see Table 2 for parameter estimates and Figure 2 for a graphical depiction). No effect was found for the collaborative condition, however. An additional fixed effect was yielded for the pretest performance ($p < .001$), suggesting that higher pretest scores were predictive of better performance in the posttest. A follow-up analysis comparing competitive and collaborative play suggested that there was no difference in posttest scores between the conditions ($p = .14$).

Furthermore, there was a significant random effect suggesting that posttest scores varied between pairs, $\chi^2(38) = 65.97$, $p = .004$. The intraclass correlation coefficient ($ICC = .40$) suggested that 40% of the variance in posttest scores could be explained by the variability between pairs, whereas 60% could be explained by the variability between players.

Our second analysis of game performance examined the number of times players received "Game Over" messages in the experimental play session, with higher numbers indicating poorer performance. In order to examine the effect mode of play had on problem-solving strategy use, we ran an HLM with number of "Game Over" messages received as the dependent variable. Dummy-coded variables for the mode-of-play conditions were included in the model along with pretest number of "Game Over" messages and prior experience with video game controllers as covariates. The equations took the following forms:

$$\text{Level 1: } Y_{ij}(\text{game over messages}) = \beta_{0j} + \beta_{1j}(\text{competition})$$

$$+ \beta_{2j}(\text{collaboration}) + \beta_{3j}(\text{pretest game over messages})$$

$$+ \beta_{4j}(\text{controller experience}) + r_{ij}.$$

Table 2
Estimates for Game Performance: Number of Correct Solutions During the Posttest

Fixed effects	Coefficient	SE
Intercept	6.61**	2.52
Competitive play	4.20*	1.61
Collaborative play	1.81	1.45
Pretest problems solved	.63***	.12
Prior controller experience	-.45	.47
Random effects		Variance
Pairs intercept		8.13**

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

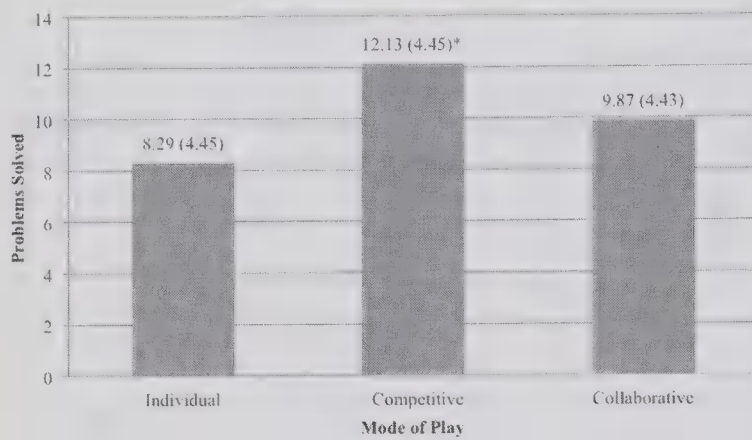


Figure 2. Adjusted means for number of problems solved in the posttest game play by condition. Values atop each bar represent means (and standard deviations). *P* value reflects comparison with the individual play condition. * $p \leq .05$.

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \beta_{1j} = \gamma_{10}, \beta_{2j} = \gamma_{20}, \beta_{3j} = \gamma_{30}, \beta_{4j} = \gamma_{40}.$$

The analysis yielded a statistically significant fixed main effect for collaborative play ($p = .004$), suggesting that players in that condition had a higher rate of inefficient problem-solving strategy use than those in the individual play condition. No such effect was found for the competitive condition in relation to the individual play condition (see Table 3 for parameter estimates and Figure 3 for a graphical depiction). A follow-up analysis comparing the competitive and collaborative conditions suggested that there was no difference in the receipt of "Game Over" messages between the groups ($p = .11$).

Furthermore, there was a significant random effect, $\chi^2(31) = 817.87$, $p < .001$, suggesting that the number of "Game Over" messages received varied between pairs. The ICC (ICC = .96) suggested that 96% of the variance in "Game Over" messages received was attributable to variability between pairs, whereas only 4% was attributable to variability between individual players.

Achievement Goal Orientations

Our next set of analyses examined the effect mode of play had on participants' adoption of achievement goal orientations during game play. In each of the three analyses, the achievement goal orientation score was entered as the dependent variable in an HLM

Table 3

Estimates for Game Performance: Number of "Game Over" Messages Received During the Experimental Trial

Fixed effects	Coefficient	SE
Intercept	.76	.59
Competitive play	1.35	.87
Collaborative play	3.53**	1.09
Pretest "Game Over" messages	.40	.23
Prior controller experience	-.02	.12
Random effects		Variance
Pairs intercept		6.13***

** $p \leq .01$. *** $p \leq .001$.

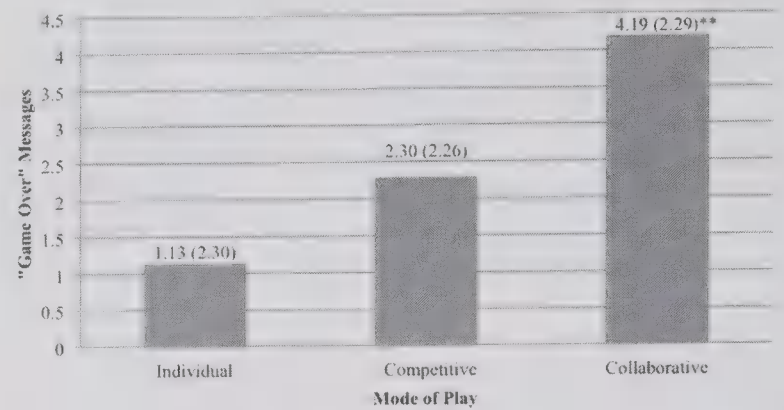


Figure 3. Adjusted means for number of "Game Over" messages received in the experimental play session by condition. Values atop each bar represent means (and standard deviations). *P* value reflects comparison with the individual play condition. ** $p \leq .01$.

along with mode-of-play condition dummy codes as predictors. The equations took the following forms:

$$\text{Level 1: } Y_{ij}(\text{goal orientation}) = \beta_{0j} + \beta_{1j}(\text{competition}) + \beta_{2j}(\text{collaboration}) + r_{ij}.$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \beta_{1j} = \gamma_{10}, \beta_{2j} = \gamma_{20}.$$

The analysis for mastery goal orientation scores yielded significant fixed main effects for both competitive ($p = .01$) and collaborative ($p = .04$) conditions, suggesting that both conditions invoked a stronger mastery goal orientation than did playing the game individually (see Figure 4 for a graphical depiction). A follow-up analysis, however, showed that mastery goal orientation strength did not differ between the competitive and collaborative groups ($p = .28$). Furthermore, there was no significant random effect, $\chi^2(38) = 38.19$, $p = .46$, suggesting that the strength of mastery goal orientations was not attributable to the variability between pairs.

For the performance-approach ($M_{\text{Ind}} = 3.99$, $SD_{\text{Ind}} = 1.32$; $M_{\text{Comp}} = 4.67$, $SD_{\text{Comp}} = 1.81$; $M_{\text{Coll}} = 3.99$, $SD_{\text{Coll}} = 1.55$) and performance-avoidance analyses ($M_{\text{Ind}} = 3.94$, $SD_{\text{Ind}} = 1.47$; $M_{\text{Comp}} = 4.32$, $SD_{\text{Comp}} = 1.77$; $M_{\text{Coll}} = 4.09$, $SD_{\text{Coll}} = 1.49$), no

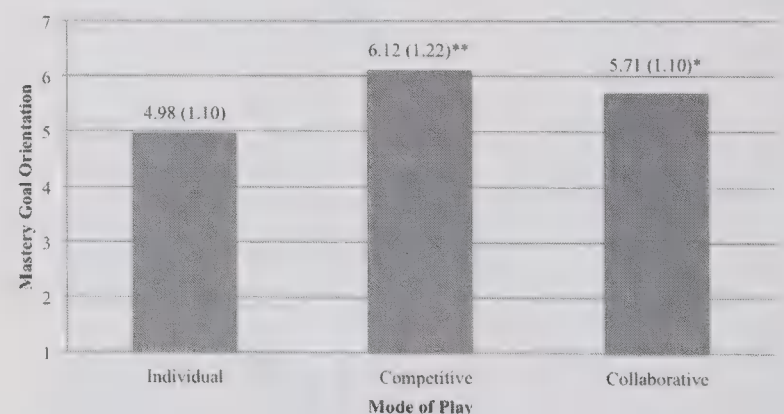


Figure 4. Mean adoption of mastery goal orientation by condition. Values atop each bar represent means (and standard deviations). *P* value reflects comparison with the individual play condition. * $p \leq .05$. ** $p \leq .01$.

significant effects were found. See Tables 4, 5, and 6 for parameter estimates for the three goal orientation models.

Situational Interest

Our next analysis examined the extent to which mode of play invoked situational interest in players during game play. Situational interest scores were entered in an HLM as the dependent variable along with the two mode-of-play condition dummy variables as predictors. The equations took the following forms:

$$\text{Level 1: } Y_{ij}(\text{situational interest}) = \beta_{0j} + \beta_{1j}(\text{competition}) \\ + \beta_{2j}(\text{collaboration}) + r_{ij}.$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \beta_{1j} = \gamma_{10}, \beta_{2j} = \gamma_{20}.$$

The analysis yielded statistically significant fixed main effects for both competitive ($p = .04$) and collaborative play conditions ($p = .01$; see Table 7 for parameter estimates and Figure 5 for a graphical depiction). These results suggest that playing in either competition or collaboration with another player made the game more exciting and personally relevant, as measured by the Situational Interest scale, than when playing it alone. A follow-up analysis comparing competitive and collaborative modes showed that they did not differ with respect to situational interest ($p = .59$). Furthermore, no random effect was yielded, $\chi^2(40) = 44.35$, $p = .29$, demonstrating that the variance in situational interest was not explained by the variability between pairs.

Enjoyment

We next examined the effect of mode of play on players' enjoyment of the game. Enjoyment scores were entered into the HLM as the dependent variable along with the dummy variables reflecting the mode-of-play conditions as predictors. The equations took the following forms:

$$\text{Level 1: } Y_{ij}(\text{enjoyment}) = \beta_{0j} + \beta_{1j}(\text{competition}) \\ + \beta_{2j}(\text{collaboration}) + r_{ij}.$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \beta_{1j} = \gamma_{10}, \beta_{2j} = \gamma_{20}.$$

The analysis yielded statistically significant fixed main effects of competitive ($p = .03$) and collaborative ($p < .001$) play on game enjoyment as compared with the individual play condition. These results suggest that playing the game alone was significantly less enjoyable than playing it either competitively or collaboratively (Table 8 lists the parameter estimates for the model, and

Table 4
Estimates for Mastery Goal Orientation

Fixed effects	Coefficient	SE
Intercept	4.98***	.27
Competitive play	1.14**	.40
Collaborative play	.74*	.33
Random effects		Variance
Pairs intercept		.01

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

Table 5
Estimates for Performance-Approach Goal Orientation

Fixed effects	Coefficient	SE
Intercept	3.99***	.32
Competitive play	.61	.51
Collaborative play	.04	.51
Random effects		Variance
Pairs intercept		.58

*** $p \leq .001$.

Figure 6 provides a graphical depiction). A follow-up analysis comparing the competitive and collaborative groups demonstrated that they did not differ with regard to their enjoyment of the game ($p = .38$).

Finally, there was a significant random effect suggesting that enjoyment of the game varied between pairs, $\chi^2(40) = 68.60$, $p = .003$. The ICC (ICC = .36) suggested that 36% of the variance in game enjoyment was explained by the variability between pairs, whereas 64% was explained by the variability between individual players.

Future Game Intentions

Two indicators of players' future intentions with regard to the game were examined. The first analysis examined the reported likelihood participants would play the game again. Reengagement intentions were entered into the hierarchical model as the dependent variable along with mode-of-play condition dummy codes. The equations took the following forms:

$$\text{Level 1: } Y_{ij}(\text{reengagement intentions}) = \beta_{0j} + \beta_{1j}(\text{competition}) \\ + \beta_{2j}(\text{collaboration}) + r_{ij}.$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \beta_{1j} = \gamma_{10}, \beta_{2j} = \gamma_{20}.$$

The analysis resulted in a statistically significant fixed main effect of the collaborative condition ($p = .03$), such that players in that condition reported a higher likelihood of playing the game again than those who played the game individually. No such effect was found for the competitive condition, however. Those participants' intentions to reengage in the game were no different than those in the individual play group. See Table 9 for the model parameter estimates and Figure 7 for a graphical depiction. A follow-up analysis further suggested that intentions to play the game again were no different for those in the competitive and collaborative play conditions ($p = .57$). Furthermore, there was no

Table 6
Estimates for Performance-Avoidance Goal Orientation

Fixed effects	Coefficient	SE
Intercept	3.94***	.36
Competitive play	.37	.59
Collaborative play	.15	.45
Random effects		Variance
Pairs intercept		.29

*** $p \leq .001$.

Table 7
Estimates for Situational Interest

Fixed effects	Coefficient	SE
Intercept	4.92***	.24
Competitive play	.74*	.33
Collaborative play	.90**	.31
Random effects		Variance
Pairs intercept		.05

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

significant random effect, $\chi^2(39) = 36.93, p > .50$, suggesting that the variance in intentions to reengage in the game was not due to variability between pairs.

The second future intention examined was players' intention to recommend the game to a friend or teacher. Therefore, recommendation intentions were added to the hierarchical model as a dependent variable along with the mode-of-play condition dummy variables. The equations took the following forms:

$$\text{Level 1: } Y_{ij}(\text{recommendation intentions}) = \beta_{0j} + \beta_{1j}(\text{competition}) + \beta_{2j}(\text{collaboration}) + r_{ij}.$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \beta_{1j} = \gamma_{10}, \beta_{2j} = \gamma_{20}.$$

The analysis yielded a statistically significant fixed effect for the collaborative condition ($p = .01$), but not the competitive condition. These results suggest that playing the game collaboratively led participants to report a stronger intention to recommend the game to someone else than those who played the game individually (see Table 10 for parameter estimates and Figure 8 for a graphical depiction). A follow-up analysis comparing the competitive and collaborative conditions yielded a null result ($p = .43$), demonstrating that intentions to recommend the game did not differ between the two groups. Furthermore, there was no significant random effect, $\chi^2(39) = 40.24, p = .42$, suggesting that recommendation intentions did not vary between pairs.

Math Fluency

Before investigating the effect of game play condition on math fluency, we first examined whether there was an overall change

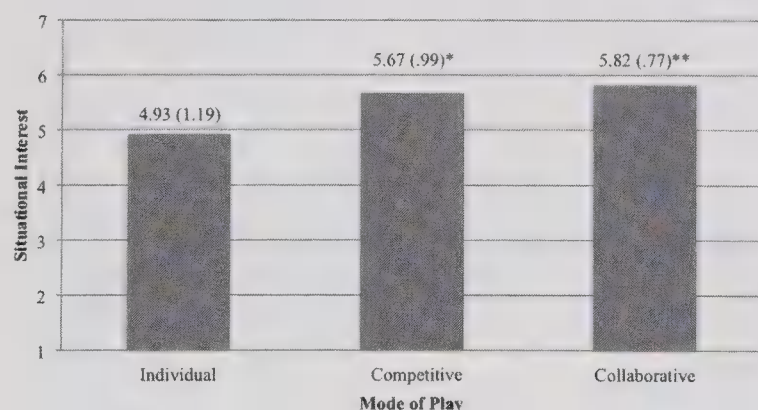


Figure 5. Mean situational interest scores by condition. Values atop each bar represent means (and standard deviations). P value reflects comparison with the individual play condition. * $p \leq .05$. ** $p \leq .01$.

Table 8
Estimates for Game Enjoyment

Fixed effects	Coefficient	SE
Intercept	7.04***	.37
Competitive play	1.33*	.53
Collaborative play	1.82***	.44
Random effects		Variance
Pairs intercept		.76**

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

from pre- to posttest fluency scores (see Table 11). A paired t test comparing pre- and posttest fluency scores was conducted; however, one participant did not complete the posttest and was thus omitted from the analysis. Results suggested that posttest fluency scores ($M = 70.42, SD = 26.67$) were statistically significantly higher than pretest scores ($M = 66.86, SD = 26.42$), $t(56) = -2.59, p = .01$. Therefore, players increased their math fluency from pre- to posttest.

Next, we analyzed posttest math fluency scores to investigate the effect of condition. Dummy-coded modes of play were entered into the model along with pretest fluency scores as a covariate. The equations took the following forms:

$$\text{Level 1: } Y_{ij}(\text{posttest fluency scores}) = \beta_{0j} + \beta_{1j}(\text{competition}) + \beta_{2j}(\text{collaboration}) + \beta_{3j}(\text{pretest fluency scores}) + r_{ij}.$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j}, \beta_{1j} = \gamma_{10}, \beta_{2j} = \gamma_{20}, \beta_{3j} = \gamma_{30}.$$

Although the effect of pretest fluency scores was found to be significant ($p < .001$), indicating a positive relation with posttest scores, the analysis indicated no effect of collaborative play or competitive play on posttest math fluency scores. The null result suggests that there were no differences in fluency scores between the individual ($M = 65.63, SD = 15.21$), competitive ($M = 78.68, SD = 36.47$), and collaborative ($M = 66.77, SD = 22.30$) game play conditions. The effect of the grouping variable, pairs, on posttest math fluency scores was found to be not statistically significant. This indicates that none of the variance in posttest math fluency scores is attributable to the pairings after accounting for variability from pretest fluency scores. Furthermore, there was

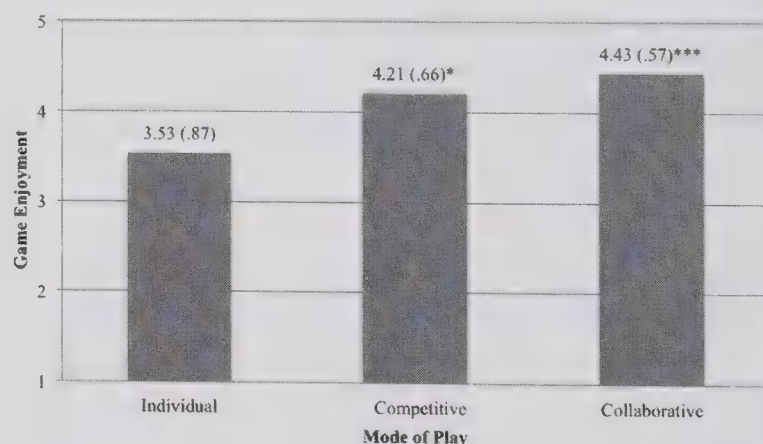


Figure 6. Mean game enjoyment scores by condition. Values atop each bar represent means (and standard deviations). P value reflects comparison with the individual play condition. * $p \leq .05$. *** $p \leq .001$.

Table 9
Estimates for Future Game Intentions: Reengagement

Fixed effects	Coefficient	SE
Intercept	3.58***	.22
Competitive play	.48	.32
Collaborative play	.63*	.26
Random effects		Variance
Pairs intercept		.00

* $p \leq .05$. *** $p \leq .001$.

no significant random effect, $\chi^2(39) = 33.01$, $p > .50$, suggesting that math fluency did not vary between pairs.

Discussion

The goal of the present research was to investigate the learning, performance, and motivational outcomes associated with playing an educational math game either competitively or collaboratively as compared with individually. With a few exceptions, our predictions were confirmed.

Two analyses were conducted to assess the affect of mode of play on within-game learning and performance. The first analysis examined the number of problems solved in the posttest, which showed that, in comparison to individual play, performance was better for competitive, but not collaborative play. Playing competitively may have aided in the development of arithmetic skills such that players were able to solve more problems during the within-game posttest. Our second analysis examined the efficiency of problem-solving strategies used by learners during the experimental session, operationalized as the number of "Game Over" screens received by the player, which found that collaborative play resulted in worse performance than individual play. There was no difference, however, in performance between competitive and individual modes of play.

There may be different explanations for these results. One possibility is that our findings may be specific to the game used in the present study. Indeed, collaboration has been shown to be beneficial for motivation and learning under numerous circumstances (e.g., Deutsch & Krauss, 1960; Hänze & Berger, 2007; Nichols, 1996; Nichols & Miller, 1994; Sharan & Shaulov, 1990;

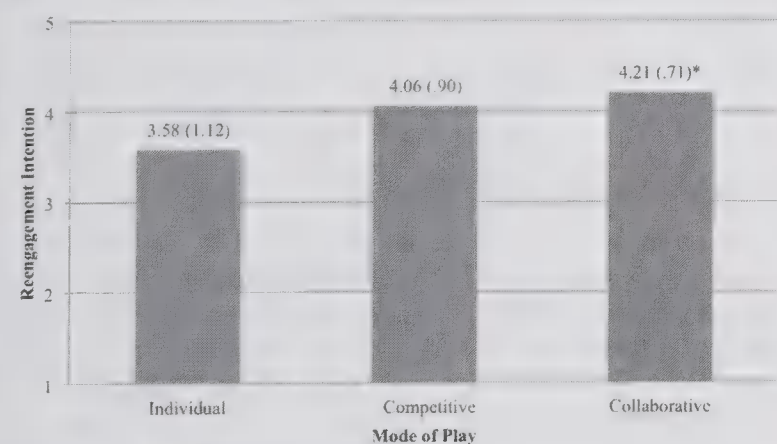


Figure 7. Mean intent to play game again by condition. Values atop each bar represent means (and standard deviations). P value reflects comparison with the individual play condition. * $p \leq .05$.

Table 10
Estimates for Future Game Intentions: Recommendation

Fixed effects	Coefficient	SE
Intercept	3.42***	.27
Competitive play	.64	.39
Collaborative play	.89**	.31
Random effects		Variance
Pairs intercept		.00

** $p \leq .01$. *** $p \leq .001$.

Slavin, 1988). *FactorReactor*, however, requires players in the collaborative mode to communicate with each other, negotiating which strategy to select, and who will execute which move. For the automation of arithmetic fluency, these particular tasks may be best suited for modes in which players are in sole control of their game space, as they were in the individual and competitive modes of the present study. This is in line with findings by Mullins et al. (2011), who found that the mutual elaborations and explanations were beneficial for conceptual knowledge, but not for skill development. Another possible explanation is that the relatively short game play penalized players for their collaborative meaning-making and exploration, which was reflected in fewer problems solved, and more inefficient strategies explored, than individual play. In a longer game play, this initial exploration may have eventually resulted in better performance than individual or competitive play, which should be investigated in future research.

It should also be noted that it is uncertain whether within-game learning occurred because players in the competitive condition had improved their math fluency or whether there were other explanations. For example, competitive players may have increased their fluency of the game mechanics relative to those in other modes of play, or honed their strategies more effectively. In other words, it is possible that they improved their game-playing skills rather than their arithmetic skills. Future research will need to investigate these possible sources of increased fluency.

Another set of analyses examined out-of-game learning, which was assessed using timed paper-and-pencil tests before and after participants played the game. It was found that players' math fluency scores had improved overall. Without additional data,

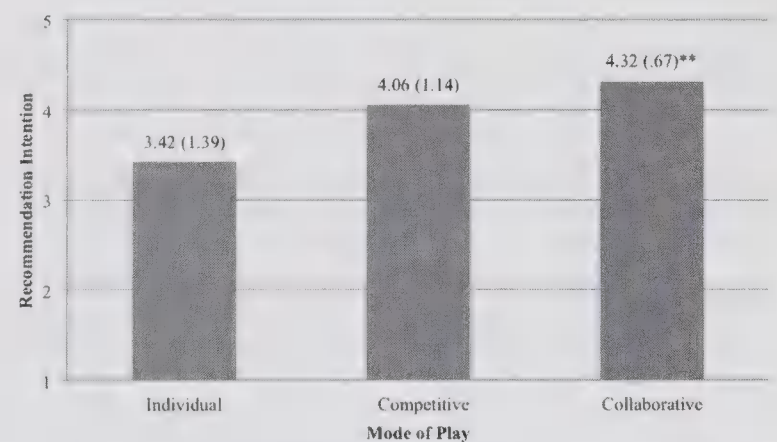


Figure 8. Mean intent to recommend game by condition. Values atop each bar represent means (and standard deviations). P value reflects comparison with the individual play condition. ** $p \leq .01$.

Table 11
Estimates for Math Fluency

Fixed effects	Coefficient	SE
Intercept	6.14	4.31
Competitive play	-2.59	3.39
Collaborative play	-.39	3.25
Pretest problems solved	.98***	.06
Random effects		Variance
Pairs intercept		.17

*** $p \leq .001$.

however, we cannot necessarily claim that the increase in scores was due to playing the game rather than reflecting a test-taking effect. Contrary to our predictions, a second analysis showed that posttest scores did not vary by condition. Although it is possible that the skills acquired during game play do not transfer to out-of-game measures of learning, the relatively short duration of game play may have not been sufficient for the transfer to occur.

A series of analyses were also conducted to examine the effect of mode of play on multiple indicators of motivation. In comparison to individual play, competitive and collaborative play resulted in the strongest mastery goal orientation, which is associated with highly adaptive patterns of motivation and learning (Ames, 1992; Midgley et al., 2001). This finding suggests that these modes of play may impact students' learning-related goals to focus more on learning the subject matter, improving, and finding the most optimal strategies, and less on normative comparisons with other students or validating their abilities. This notion is further supported by the fact that we found that the competitive and collaborative modes of play did not differ from individual play in their invocation of performance-approach and performance-avoidance goal orientations. This null finding may have stemmed from the way in which participants experienced the competitive mode. Although performance goals are concerned with outperforming others, it is in the service of demonstrating normative ability (e.g., Grant & Dweck, 2003; Urda & Mestas, 2006). Indeed, performance goals and competition are different constructs. Playing in competition with another student may not be sufficient to invoke concerns about normative ability. If a student were to play against all of his or her classmates and their scores were made available to each other, however, a concern for normative performance may be elicited, along with a performance goal. In other words, although competition may play a role in the invocation of performance goals, such that there exists a desire to outperform others, our data suggest that the way in which competition was operationalized in *FactorReactor* was not sufficient to invoke performance goal orientations. Given the properties of the game we used, our results suggest that in the context of a learning game, competition with only one other player, rather than all other classmates, may be an effective means of invoking a mastery goal orientation without the negative outcomes associated with the invocation of performance goal orientations.

Our results also demonstrated that competitive and collaborative play increased situational interest and game enjoyment in relation to individual play. That these constructs were augmented has particularly important implications for the use of these modes of play in educational games. First, students are more likely to engage

in a task they perceive to be enjoyable (Salen & Zimmerman, 2003), thereby increasing their exposure to the educational content. Second, the invocation of situational interest suggests that the effect of the game reaches beyond mere enjoyment. Relative to the individual play condition, players in the competitive and collaborative conditions experienced the game as personally involving and that the content of the game was valuable and personally relevant. This increase in situational interest lays a foundation on which a more internalized and enduring interest, individual interest, is built.

Additionally, collaborative play increased participants' intention to play the game again and to recommend the game to another. This supports the notion that games not only engage students in particular learning activities and content but also increases the likelihood of reengagement over time, in and out of classroom (Gee, 2007; Squire, 2003). It also suggests that they may foster the development of a more internalized individual interest that intrinsically guides students' future learning endeavors, both alone and assisted by an instructor (Bergin, 1999; Deci, Vallerand, Pelletier, & Ryan, 1991).

It should be noted that our indicators of motivation were generally assessed in terms of the game itself. Therefore, it is possible that our results reflect motivational responses to the game rather than arithmetic. The intention of educational games, however, is to provide a context that engages learners and motivates them to reengage over time. Furthermore, the effectiveness of these games is attributable, in part, to their ability to reengage learners. For example, a student who enjoys a math game may play it frequently, resulting in increased exposure to and practice with mathematical operations. Even so, a number of the items used to assess motivation referred specifically to the learning content of the game, as with our assessment of situational interest, which likewise resulted in our predicted effects.

Taken together, our results suggest that there are benefits and costs associated with particular modes of play. Although the competitive and collaborative modes elicited the strongest motivation and interest, and increased the degree to which mastery goal orientations were adopted, the collaborative condition resulted in the highest frequency of inefficient strategy use, yet led to more positive attitudes toward the game. More specifically, participants in the collaborative condition had to restart the most levels, suggesting that their collaborations were inefficient and error-prone, and led to the use of poor strategies as compared to those in the individual mode. Yet, collaborative play also led to greater intentions to play the game again, suggesting that, over time, this negative effect could be resolved.

Limitations and Future Research

As is the case for all empirical studies, there are some limitations to the generalizability of our findings. Most importantly, the results of this study cannot be readily generalized to all educational games. The game used in this research, *FactorReactor*, was designed for the practice and automation of arithmetic skills to increase fluency in middle-school students. There are many genres of games with features that differ significantly from this game, such as role-playing games, adventure games, augmented reality games, or first-person shooters. Because of the corresponding design differences, a different effect of mode of play might be

expected for other game genres. For example, we found collaborative play to increase the rate of adoption of inefficient and error-prone strategies during the game. We would not, however, suggest that collaboration is detrimental to performance in general. The characteristics of this particular game may not have been ideal for collaborative play within a 15-min period, which may have been alleviated by the fact that we chose, for the reasons outlined above, to keep instruction on how to collaborate to a minimum. Future research will need to investigate whether our findings can be replicated with games with similar objectives, but from other game genres. Future work should also examine the effects of other design factors, and should investigate whether the effects found in the present study are different for games that cover different kinds of knowledge, for example, whether collaborative game play would result in better learning of conceptual knowledge, as suggested by Mullins et al. (2011). We are also interested in conducting further research to explore the learning processes in individual versus collaborative and competitive game modes by collecting process data such as biometrics and eye tracking (Aleven, Rau, & Rummel, 2012).

Conclusion

The results of this study, which provide initial evidence for the effect of social context in game-based development of arithmetic fluency, have important theoretical and practical implications.

On the theoretical side, we demonstrated that although only the competitive mode of play increased within-game learning, both competitive and collaborative modes of play increased situational interest, enjoyment, and the adoption of a mastery goal orientation, compared with individual play. These results are in line with previous research in computer-supported learning of mathematics that showed that benefits of collaboration were only found for conceptual knowledge, but not found for skills acquisition (Mullins et al., 2011). Our research extended these findings by also considering the impact of a form of competition that has the benefits of increased performance while still invoking a mastery goal orientation rather than a performance goal orientation. It is especially interesting that although resulting in inefficient use of problem-solving strategies and error-prone game play, collaborative play was associated with greater enjoyment, situational interest, and intention of reengagement than individual play. These results fit within a framework of learning with media that recognizes the importance of social context and related affective variables in addition to cognitive ones (Moreno & Mayer, 2007).

On the practical side, this research provides empirical support for the potential of educational games as effective learning environments that provide incentives for students to play repeatedly over time. Our results demonstrate that game designers need to earnestly consider the differential effects of competitive and collaborative modes of a game in skill fluency development. Although both modes of social play increase situational interest and future intentions to play, only the competitive mode resulted in increases in game performance compared with individual play, whereas collaborative play resulted in the adoption of less efficient problem-solving strategies. This research also highlights that many of the outcomes of learning with gamelike environments are of an affective nature and that such affective outcomes of motivation

and interest have to be considered in addition to the cognitive learning outcomes of a game.

In summary, the research reported in this study provides empirical support for a *social context* design pattern that emphasizes competitive modes of play over collaborative and individual play for games aimed at developing arithmetic skill fluency and adopting of a mastery goal orientation, as well as increasing situational interest and enjoyment.

References

- Aleven, V., Rau, M., & Rummel, N. (2012). *Planned use of eye movement data to explore complementary strengths of individual and collaborative learning*. Proceedings of the DUET 2012 - Dual Eye-Tracking in CSCW Meeting. Retrieved from http://dualeyettracking.org/duet2012/Program_files/DUET2012_1.pdf
- Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language: Towns, buildings, construction* (Vol. 2). New York, NY: Oxford University Press.
- Ames, C. (1984). Achievement attributions and self-instructions under competitive and individualistic goal structures. *Journal of Educational Psychology*, 76, 478–487. doi:10.1037/0022-0663.76.3.478
- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84, 261–271. doi:10.1037/0022-0663.84.3.261
- Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology*, 80, 260–267. doi:10.1037/0022-0663.80.3.260
- Ames, C., & Felker, D. (1979). An examination of children's attributions and achievement-related evaluations in competitive, cooperative, and individualistic reward structures. *Journal of Educational Psychology*, 71, 413–420. doi:10.1037/0022-0663.71.4.413
- Barab, S., Ingram-Goble, A., & Warren, S. (2008). Conceptual play spaces. In R. E. Ferdig (Ed.), *Handbook of research on effective electronic gaming in education* (pp. 989–1009). New York, NY: IGI Global.
- Barab, S. A., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development*, 53, 86–107. doi:10.1007/BF02504859
- Berg, K. F. (1994). *Scripted cooperation in high school mathematics: Peer interaction and achievement*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Bergin, D. (1999). Influences on classroom interest. *Educational Psychologist*, 34, 87–98. doi:10.1207/s15326985ep3402_2
- Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: Handbook I. Cognitive domain*. New York, NY: McKay.
- Chaudhuri, S., Kumar, R., Joshi, M., Terrell, E., Higgs, F., Aleven, V., & Rosé, C. P. (2008). It's not easy being green: Supporting collaborative "green design" learning. In E. Aimeur & B. Woolf (Eds.), *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems* (pp. 807–809). Berlin, Germany: Springer-Verlag.
- Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology*, 93, 43–54. doi:10.1037/0022-0663.93.1.43
- Cobb, P., Wood, T., & Yackel, E. (1991). A constructivist approach to second-grade mathematics. In E. von Glasersfeld (Ed.), *Constructivism in mathematics education* (pp. 157–176). Dordrecht, the Netherlands: Kluwer.
- Collins, A., & Halverson, R. (2009). *Rethinking education in the age of technology: The digital revolution and schooling in America*. New York, NY: Teachers College.
- Cury, F., Elliot, A. J., Da Fonseca, D., & Moller, A. C. (2006). The Social-cognitive model of achievement motivation and the 2×2

- achievement goal framework. *Journal of Personality and Social Psychology*, 90, 666–679. doi:10.1037/0022-3514.90.4.666
- Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist*, 26, 325–346.
- Deutsch, M., & Krauss, R. M. (1960). The effect of threat upon interpersonal bargaining. *Journal of Abnormal and Social Psychology*, 61, 181–189. doi:10.1037/h0042589
- Dillenbourg, P. (1999). Introduction: What do you mean by 'collaborative learning'? In P. Dillenbourg (Ed.), *Collaborative learning – cognitive and computational approaches* (pp. 1–19). Amsterdam, the Netherlands: Pergamon.
- Diziol, D., Rummel, N., Spada, H., & McLaren, B. (2007). Promoting learning in mathematics: Script support for collaborative problem solving with the Cognitive Tutor Algebra. In C. A. Chinn, G. Erkens, & S. Puntambekar (Eds.), *Mice, minds and society: Proceedings of the Computer Supported Collaborative Learning (CSCL) Conference 2007* (Vol. 8, pp. 39–41). New Brunswick, NJ: International Society of the Learning Sciences.
- Diziol, D., Walker, E., Rummel, N., & Koedinger, K. R. (2010). Using intelligent tutor technology to implement adaptive support for student collaboration. *Educational Psychology Review*, 22, 89–102.
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41, 1040–1048. doi:10.1037/0003-066X.41.10.1040
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256–273. doi:10.1037/0033-295X.95.2.256
- Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*, 34, 169–189. doi:10.1207/s15326985ep3403_3
- Elliot, A. J. (2005). A conceptual history of the achievement goal construct. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). New York, NY: Guilford Publications.
- Elliot, A. J., & Mapes, R. R. (2005). Approach-avoidance motivation and self-concept evaluation. In A. Tesser, J. V. Wood, & D. A. Stapel (Eds.), *On building, defending and regulating the self: A psychological perspective* (pp. 171–196). New York, NY: Psychology Press.
- Elliot, A. J., & McGregor, H. A. (1999). Test anxiety and the hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, 76, 628–644. doi:10.1037/0022-3514.76.4.628
- Elliot, A. J., & McGregor, H. A. (2001). A 2×2 achievement goal framework. *Journal of Personality and Social Psychology*, 80, 501–519. doi:10.1037/0022-3514.80.3.501
- Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology*, 54, 5–12. doi:10.1037/0022-3514.54.1.5
- Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction*, 12, 213–232.
- Fu, F., Wu, Y., & Ho, H. (2009). An investigation of cooperative pedagogic design for knowledge creation in web-based learning. *Computers & Education*, 53, 550–562. doi:10.1016/j.compedu.2009.01.004
- Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York, NY: Palgrave Macmillan. doi:10.1145/950566.950595
- Gee, J. P. (2007). *Good video games + Good learning*. New York, NY: Peter Lang.
- Grant, H., & Dweck, C. S. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology*, 85, 541–553. doi:10.1037/0022-3514.85.3.541
- Hänze, M., & Berger, R. (2007). Cooperative learning, motivational effects, and student characteristics: An experimental study comparing cooperative learning and direct instruction in 12th grade physics classes. *Learning and Instruction*, 17, 29–41. doi:10.1016/j.learninstruc.2006.11.004
- Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology*, 94, 638–645. doi:10.1037/0022-0663.94.3.638
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology*, 92, 316–330. doi:10.1037/0022-0663.92.2.316
- Hidi, S., & Renninger, K. A. (2006). The Four-phase model of interest development. *Educational Psychologist*, 41, 111–127. doi:10.1207/s15326985ep4102_4
- Hron, A., & Friedrich, H. F. (2003). A review of Web-based collaborative learning: Factors beyond technology. *Journal of Computer Assisted Learning*, 19, 70–79.
- Hron, A., Hesse, F. W., Cress, U., & Giovis, C. (2000). Implicit and explicit dialogue structuring in virtual learning groups. *British Journal of Educational Psychology*, 70, 53–64.
- Johnson, D. W., Johnson, R. T., Maruyama, G., Nelson, D., & Skon, L. (1981). Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta analysis. *Psychological Bulletin*, 89, 47–62. doi:10.1037/0033-2909.89.1.47
- Johnson, R. T., Johnson, D. W., & Stanne, M. B. (1986). Comparison of computer-assisted cooperative, competitive, and individualistic learning. *American Educational Research Journal*, 23, 382–392.
- Kafai, Y. (1995). *Minds in play: Computer game design as a context for children's learning*. Mahwah, NJ: Lawrence Erlbaum.
- Karabenick, S. A. (2004). Perceived achievement goal structure and college student help seeking. *Journal of Educational Psychology*, 96, 569–581. doi:10.1037/0022-0663.96.3.569
- Ke, F., & Grabowski, B. (2007). Gameplay for maths learning: Cooperative or not? *British Journal of Educational Technology*, 38, 249–259. doi:10.1111/j.1467-8535.2006.00593.x
- Kirschner, F., Paas, F., Kirschner, P. A., & Janssen, J. (2011). Differential effects of problem-solving demands on individual and collaborative learning outcomes. *Learning and Instruction*, 21, 587–599. doi:10.1016/j.learninstruc.2011.01.001
- Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology: General*, 124, 83–97. doi:10.1037/0096-3445.124.1.83
- Light, P., & Littleton, K. (1999). *Social processes in children's learning*. Cambridge, UK: Cambridge University Press.
- Linnenbrink, E. A. (2005). The dilemma of performance-approach goals: The use of multiple goal contexts to promote students' motivation and learning. *Journal of Educational Psychology*, 97, 197–213. doi:10.1037/0022-0663.97.2.197
- Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement*, 70, 647–671. doi:10.1177/0013164409355699
- Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Apollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, 66, 423–458.
- Maehr, M. L., & Zusho, A. (2009). Achievement goal theory: The past, present, and future. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 77–104). New York, NY: Routledge/Taylor & Francis Group.
- Mayo, M. J. (2007). Games for science and engineering education. *Communications of the ACM*, 50, 30–35. doi:10.1145/1272516.1272536
- McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual: Woodcock-Johnson III*. Itasca, IL: Riverside.

- Meece, J. L., Anderman, E. M., & Anderman, L. H. (2006). Classroom goal structures, student motivation, and academic achievement. *Annual Review of Psychology*, 57, 487–503. doi:10.1146/annurev.psych.56.091103.070258
- Meece, J. L., Blumenfeld, P. C., & Hoyle, R. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology*, 80, 514–523. doi:10.1037/0022-0663.80.4.514
- Mevarech, Z. R., Stern, D., & Levita, I. (1987). To cooperate or not to cooperate in CAI: That is the question. *Journal of Educational Research*, 80, 164–167.
- Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of lack of ability: An underexplored aspect of goal theory. *Journal of Educational Psychology*, 89, 710–718. doi:10.1037/0022-0663.89.4.710
- Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance-approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology*, 93, 77–86. doi:10.1037/0022-0663.93.1.77
- Midgley, C., Kaplan, A., Middleton, M., Maehr, M. L., Urdan, T., Hicks, L., . . . Roeser, R. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology*, 23, 113–131. doi:10.1006/ceps.1998.0965
- Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L. H., Freeman, K. E., . . . Urdan, T. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS)*. Ann Arbor: University of Michigan.
- Moreno, R., & Mayer, R. E. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, 19, 309–326. doi:10.1007/s10648-007-9047-2
- Mullins, D., Rummel, N., & Spada, H. (2011). Are two heads always better than one? Differential effects of collaboration on students' computer-supported learning in mathematics. *International Journal of Computer-Supported Collaborative Learning*, 6, 421–443. doi:10.1007/s11412-011-9122-z
- Nichols, J. D. (1996). The effects of cooperative learning on achievement and motivation in a high school geometry class. *Contemporary Educational Psychology*, 21, 467–476. doi:10.1006/ceps.1996.0031
- Nichols, J. D., & Miller, R. B. (1994). Cooperative learning and student motivation. *Contemporary Educational Psychology*, 19, 167–178. doi:10.1006/ceps.1994.1015
- O'Keefe, P. A., Ben-Eliyahu, A., & Linnenbrink-Garcia, L. (2013). Shaping achievement goal orientations in a mastery-structured environment and concomitant changes in related contingencies of self-worth. *Motivation & Emotion*, 37, 50–64. doi:10.1007/s11031-012-9293-6
- Pekrun, R., Elliot, A. J., & Maier, M. A. (2009). Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of Educational Psychology*, 101, 115–135. doi:10.1037/a0013383
- Piaget, J. (1932). *The moral judgment of the child*. Oxford, England: Harcourt, Brace.
- Plass, J. L., Homer, B. D., Chang, Y. K., Frye, J., Kaczetow, W., Isbister, K., & Perlin, K. (2013). Metrics to assess learning and measure learner variables in simulations and games. In M. S. El-Nasr, A. Drachen, & A. Canossa (Eds.), *Game analytics—maximizing the value of player data* (pp. 697–729). New York, NY: Springer.
- Plass, J. L., Homer, B. D., & Hayward, E. (2009). Design factors for educationally effective animations and simulations. *Journal of Computing in Higher Education*, 21, 31–61. doi:10.1007/s12528-009-9011-x
- Plass, J. L., Homer, B. D., Milne, C., Jordan, T., Kalyuga, S., Kim, M., & Lee, H. J. (2009). Design factors for effective science simulations: Representation of information. *International Journal of Gaming and Computer-Mediated Simulations*, 1, 16–35. doi:10.4018/jgcms.2009010102
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2011). *HLM 7 for Windows [Computer software]*. Skokie, IL: Scientific Software International.
- Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology*, 91, 175–189. doi:10.1037/0022-0663.91.1.175
- Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, 93, 346–362. doi:10.1037/0022-0663.93.2.346
- Salen, K., & Zimmerman, E. (2003). *Rules of play: Game design fundamentals*. Cambridge, MA: MIT Press.
- Salomon, G. (Ed.). (1993). *Distributed cognitions: Psychological and educational considerations*. Cambridge, England: Cambridge University Press.
- Scardamalia, M., & Bereiter, C. (1991). Higher levels of agency for children in knowledge building: A challenge for the design of new knowledge media. *Journal of the Learning Sciences*, 1, 37–68. doi:10.1207/s15327809jls0101_3
- Shaffer, D. W. (2008). *How computers help children learn*. New York, NY: Palgrave.
- Sharan, S., & Shaulov, A. (1990). Cooperative learning, motivation to learn, and academic achievement. In S. Sharan (Ed.), *Cooperative learning: Theory and research* (pp. 173–202). New York, NY: Praeger.
- Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science*, 9, 405–410. doi:10.1111/1467-9280.00076
- Skaalvik, E. (1997). Self-enhancing and self-defeating ego orientations: Relations with task and avoidance orientation, achievement, self-perceptions and anxiety. *Journal of Educational Psychology*, 89, 71–81. doi:10.1037/0022-0663.89.1.71
- Slavin, R. E. (1980). Cooperative learning. *Review of Educational Research*, 50, 315–342.
- Slavin, R. E. (1983). When does cooperative learning increase student achievement? *Psychological Bulletin*, 94, 429–445. doi:10.1037/0033-2909.94.3.429
- Slavin, R. E. (1988). Cooperative learning and student achievement. *Educational Leadership*, 46, 31–33.
- Slavin, R. E., Leavey, M. B., & Madden, N. A. (1984). Combining cooperative learning and individualized instruction: Effects on student mathematics achievement, attitudes, and behaviors. *Elementary School Journal*, 84, 409–422. doi:10.1086/461373
- Squire, K. (2003). Video games in education. *International Journal of Intelligent Games & Simulation*, 2, 49–62.
- Squire, K. (2005). Changing the game: What happens when video games enter the classroom. *Journal of Online Education*, 1, 1–20.
- Squire, K. (2008). Open-ended video games: A model for developing learning for the interactive age. In K. Salen (Ed.), *The ecology of games* (pp. 167–198). Cambridge, MA: MIT Press.
- Steinkuehler, C. A. (2006). Massively multiplayer online videogaming as participation in a discourse. *Mind, Culture & Activity*, 13, 38–52. doi:10.1207/s15327884mca1301_4
- Stipek, D. J., & Kowalski, P. (1989). Learned helplessness in task-orienting versus performance orienting testing conditions. *Journal of Educational Psychology*, 81, 384–391. doi:10.1037/0022-0663.81.3.384
- Strommen, E. F. (1993). “Does yours eat leaves?” cooperative learning in an educational software task. *Journal of Computing in Childhood Education*, 4, 45–56.
- Tas, Y., & Tekkaya, C. (2010). Personal and contextual factors associated 2025 with students' cheating in science. *The Journal of Experimental Education*, 78, 440–463. doi:10.1080/00220970903548046
- Tronsky, L. N. (2005). Strategy use, the development of automaticity, and working memory involvement in complex multiplication. *Memory & Cognition*, 33, 927–940. doi:10.3758/BF03193086

- Um, E., Plass, J. L., Hayward, E. O., & Homer, B. D. (2012). Emotional design in multimedia learning. *Journal of Educational Psychology, 104*, 485–498. doi:10.1037/a0026609
- Urdan, T., & Mestas, M. (2006). The goals behind performance goals. *Journal of Educational Psychology, 98*, 354–365. doi:10.1037/0022-0663.98.2.354
- van Bruggen, J. M., Kirschner, P. A., & Jochems, W. (2002). External representation of argumentation in CSCL and the management of cognitive load. *Learning and Instruction, 12*, 121–138.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Yackel, E., Cobb, P., & Wood, T. (1991). Small-group interactions as a source of learning opportunities in second-grade mathematics. *Journal for Research in Mathematics Education, 22*, 390–408. doi:10.2307/749187

Received December 15, 2011

Revision received January 28, 2013

Accepted March 4, 2013 ■

New Editors Appointed, 2015–2020

The Publications and Communications Board of the American Psychological Association announces the appointment of 6 new editors for 6-year terms beginning in 2015. As of January 1, 2014, manuscripts should be directed as follows:

- *Behavioral Neuroscience* (<http://www.apa.org/pubs/journals/bne/>), **Rebecca Burwell, PhD**, Brown University
- *Journal of Applied Psychology* (<http://www.apa.org/pubs/journals/apl/>), **Gilad Chen, PhD**, University of Maryland
- *Journal of Educational Psychology* (<http://www.apa.org/pubs/journals/edu/>), **Steve Graham, EdD**, Arizona State University
- *JPSP: Interpersonal Relations and Group Processes* (<http://www.apa.org/pubs/journals/psp/>), **Kerry Kawakami, PhD**, York University, Toronto, Ontario, Canada
- *Psychological Bulletin* (<http://www.apa.org/pubs/journals/bul/>), **Dolores Albarracín, PhD**, University of Pennsylvania
- *Psychology of Addictive Behaviors* (<http://www.apa.org/pubs/journals/adb/>), **Nancy M. Petry, PhD**, University of Connecticut School of Medicine

Electronic manuscript submission: As of January 1, 2014, manuscripts should be submitted electronically to the new editors via the journal's Manuscript Submission Portal (see the website listed above with each journal title).

Current editors Mark Blumberg, PhD, Steve Kozlowski, PhD, Arthur Graesser, PhD, Jeffry Simpson, PhD, Stephen Hinshaw, PhD, and Stephen Maisto, PhD, will receive and consider new manuscripts through December 31, 2013.

Guiding Learners Through Technology-Based Instruction: The Effects of Adaptive Guidance Design and Individual Differences on Learning Over Time

Adam M. Kanar
Brock University

Bradford S. Bell
Cornell University

Adaptive guidance is an instructional intervention that helps learners to make use of the control inherent in technology-based instruction. The present research investigated the interactive effects of guidance design (i.e., framing of guidance information) and individual differences (i.e., pretraining motivation and ability) on learning basic and strategic task skills over time. One hundred thirty participants were randomly assigned to 1 of 2 types of adaptive guidance (autonomy supportive, controlling) or a no-guidance condition while learning to perform a complex simulation task over 9 consecutive trials. Results indicated that participants receiving controlling guidance acquired strategic task skills at a faster rate than participants receiving autonomy-supportive guidance or no guidance. The design of adaptive guidance also moderated the effects of pretraining motivation and cognitive ability on learners' acquisition of basic and strategic task skills. Specifically, autonomy-supportive guidance enhanced the positive effects of pretraining motivation on the acquisition of basic task skills, and controlling guidance enhanced the positive effects of cognitive ability on the acquisition of strategic task skills. Implications for research and practice are discussed.

Keywords: learning, technology, guidance, individual differences, performance

Over the past decade, a number of different forces, including technological advances, economic pressures, and globalization, have spurred significant growth in technology-based instruction in both higher education and corporate settings. For instance, the National Center for Education Statistics estimates that from 2000 to 2008, the percentage of undergraduates enrolled in at least one distance education course grew from 8% to 20% (Radford, 2011). Similarly, the American Society for Training and Development estimates that the percentage of learning delivered through technology in work organizations has increased from 8.8% in 2000 to 38.5% in 2011 (Miller, 2012; Van Buren & Erskine, 2002).

One important implication of this trend in learning delivery is that technology-based instruction often provides learners with significant control over different aspects (e.g., content, sequence, pace) of their learning (DeRouin, Fritzsche, & Salas, 2004). Kraiger and Jerden (2007), for example, noted that many modern forms of technology-based instruction follow a learner-centered format in which the software serves as a learning portal, and individuals must make choices about both what and how to learn.

When compared with conditions in which instructional software controls most or all of the learning decisions (i.e., program control), learner control often has a positive, albeit small, effect on student outcomes (Kraiger & Jerden, 2007; Reeves, 1993). Yet, researchers have also noted that instruction that offers high levels of learner control often proves ineffective because learners experience resource depletion, fail to come into contact with important information, and make poor learning decisions (Brown, 2001; Kirschner, Sweller, & Clark, 2006; Mayer, 2004).

These findings highlight the need for instructional strategies that can assist learners in making effective use of the control offered by technology-based instruction. One approach that has been examined involves supplementing learner control with adaptive guidance, which provides learners with diagnostic and interpretive information designed to help them make more effective learning decisions (Bell & Kozlowski, 2002). Although research has shown that adaptive guidance leads to better learning outcomes than either total learner or program control (e.g., Bell & Kozlowski, 2002; Corbalan, Kester, & van Merriënboer, 2008), the issue of how adaptive guidance should be designed to optimize student learning in technology-based instruction remains largely unexplored.

One instructional design feature that may have an important impact on student achievement is the framing of guidance information. Prior research has demonstrated that how learning instructions and activities are framed can have a significant impact on learning (e.g., Kozlowski & Bell, 2006; Rawsthorne & Elliot, 1999). For instance, drawing on self-determination theory (SDT), investigators have shown that learning contexts that are framed as autonomy supportive lead to higher levels of motivation and

This article was published Online First September 9, 2013.

Adam M. Kanar, Goodman School of Business, Brock University, St. Catharines, Ontario, Canada; Bradford S. Bell, ILR School, Cornell University.

This research was supported in part by a grant from the Center for Advanced Human Resource Studies at Cornell University.

Correspondence concerning this article should be addressed to Adam M. Kanar, Goodman School of Business, Brock University, 415 Taro Hall, St. Catharines, Ontario, Canada L2N 6P6. E-mail: akanar@brocku.ca

learning than contexts that are framed as controlling (e.g., Black & Deci, 2000; Vansteenkiste, Simons, Lens, Sheldon, & Deci, 2004). These findings suggest that guidance information should be framed so as to minimize perceptions of external control and emphasize learners' autonomy and freedom. Resource allocation theories of self-regulation (e.g., Kanfer & Ackerman, 1989), however, suggest that providing greater autonomy and choice may deplete learners' cognitive resources and impede skill acquisition, particularly in learning contexts that impose substantial demands on attentional resources. Thus, guidance that is framed as more controlling and restrictive may reduce the burden on learners, allow them to direct more of their attentional resources to learning, and increase the likelihood that learners' come into contact with important learning content (Mayer, 2004).

In the current study, we explore these different perspectives through an examination of the effects of two forms of adaptive guidance—autonomy supportive and controlling—on learning during a complex simulation-based training program. This effort advances the existing literature in at least three ways. First, using SDT and resource allocation theory, we propose that the effects of different adaptive guidance designs may vary across different learning outcomes. To test this prediction, we examine the effects of autonomy-supportive and controlling guidance on multiple indicators of learning, namely, the acquisition of basic and strategic task skills. Second, recent studies suggest that individual differences often moderate the effects of interventions designed to improve learning during technology-based instruction, such that a specific intervention will be more effective for some learners than others (e.g., Sitzmann, Bell, Kraiger, & Kanar, 2009). Building on and extending these findings, we examine how different forms of

guidance interact with individual differences related to effort (i.e., pretraining motivation) and resource availability (i.e., cognitive ability) to influence learning. Finally, we use a longitudinal design and latent growth modeling to examine the effects of the two forms of adaptive guidance over time. Whereas most research has treated the effects of guidance as static, our longitudinal approach examines the impact of the different types of adaptive guidance on individuals' learning trajectories over the course of instruction, which provides further insight into how different forms of guidance influence the acquisition of different types of task skills. The conceptual model examined in this research is presented in Figure 1. In the following sections, we discuss the theory that underlies the relationships outlined in the model.

Adaptive Guidance

Although there is some evidence that learner control can enhance student motivation and satisfaction (e.g., Reeves, 1993), research suggests that individuals often do not make effective use of the control they are given over their instruction (Steinberg, 1977, 1989). Learners frequently misinterpret feedback and are poor judges of their performance and progress, which can lead to poor learning choices and misdirected effort. Brown (2001), for example, studied learner choices during online instruction and found that learners commonly skipped critical practice opportunities, and some spent less than 50% of the available time in the course. He concluded, "Results suggest that, despite the appeal of computer-based training as a way to make learning more efficient, employees may not use control over their learning wisely" (p. 290). Mayer (2004) leveled similar criticisms against discovery

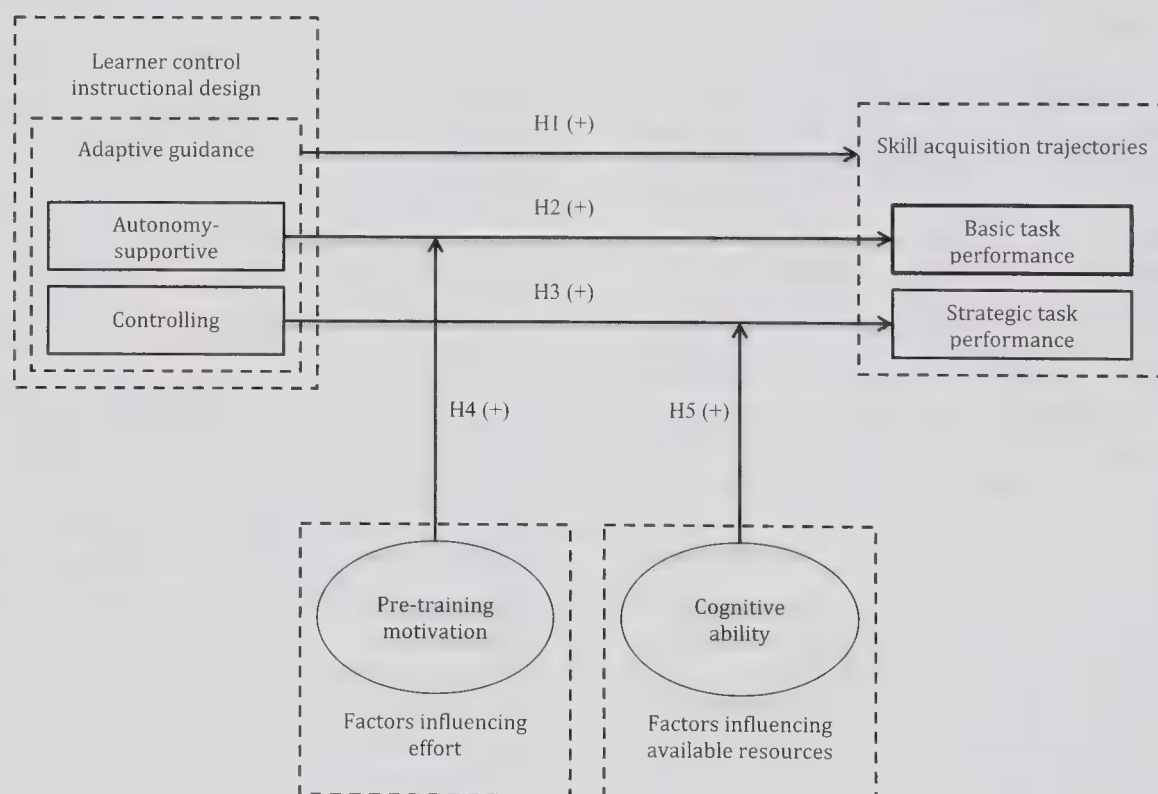


Figure 1. Conceptual model of predicted relationships between adaptive-guidance design, individual-difference factors, and performance trajectories. H1, H2, H3, H4, and H5 = Hypothesis 1, Hypothesis 2, Hypothesis 3, Hypothesis 4, and Hypothesis 5.

learning, in which students are free to work in the learning environment with little or no guidance. He reviewed research that compared pure and guided discovery methods and concluded that guided methods help ensure that students come into contact with to-be-learned material and better support the cognitive processes necessary for constructivist learning. Finally, Kirschner et al. (2006) argued that unguided environments create a heavy working memory load that is detrimental to learning.

Although guided instruction can take many forms (cf. Kirschner et al., 2006), in the current study we focus on adaptive guidance, which was designed for more complex learning environments that leverage technology (Bell & Kozlowski, 2002). Adaptive guidance was developed on the basis of a foundation provided by learner control research (e.g., Tennyson, 1980; Tennyson & Buttrey, 1980), but was also designed to extend to more complex learning domains that require learners to acquire not only basic but also strategic task skills. Basic task skills involve a trainees' ability to perform fundamental task operations that must be learned in order to develop more advanced skills (Bell & Kozlowski, 2002). Individuals use their declarative knowledge (e.g., knowledge of facts) and procedural knowledge (e.g., knowledge of rules) when performing basic skills. Through practice and experience, declarative knowledge is compiled or proceduralized, which allows trainees to execute basic operations more quickly and with fewer errors (Anderson, 1983). Strategic skills involve carrying out more difficult operations that require trainees to understand the underlying complexities of a task and integrate task concepts. In addition, trainees must develop contextual knowledge that informs why, when, and where to apply their strategic skills (Ford & Kraiger, 1995). Thus, strategic performance involves selectively retrieving and integrating specific knowledge from one's knowledge base and applying the resulting constructions to varying task contingencies (Tennyson & Breuer, 2002). In environments that require both basic and strategic skills, learning is a function of not only effort (e.g., time on task) but also the quality of study and practice activities. Thus, adaptive guidance uses learners' past performance to provide evaluative and diagnostic information that assists them in judging their progress toward task mastery, which should influence the amount of effort they invest in learning. In addition, it provides individualized suggestions for what learners should study and practice, based on progress, which should influence the allocation of attention and lead to better learning choices.

Bell and Kozlowski (2002) showed that adaptive guidance helps learners to make better learning decisions in a learner-control environment. Learners who received guidance studied and practiced training material in a more appropriate sequence than those who received no guidance. Guidance also had a positive effect on trainees' self-efficacy early in training, when learning is most challenging and errors are common. The result was that learners who received adaptive guidance exhibited higher levels of basic and strategic knowledge and performance and were better able to transfer their skills than those who were given learner control without guidance (Bell & Kozlowski, 2002). Accordingly, we expect that learners receiving adaptive guidance will exhibit greater positive change in their performance relative to those in a no-guidance condition.

Hypothesis 1: Participants who receive adaptive guidance will exhibit more positive change in basic and strategic performance skills than participants who do not receive guidance.

Autonomy-Supportive and Controlling Guidance

The issue of how adaptive guidance should be designed to optimize student learning in technology-based instruction has received limited research attention. Adaptive guidance seeks to provide the direction learners need to avoid making poor learning decisions while retaining the motivational benefits of autonomy. SDT (for a review, see Ryan & Deci, 2000) is a theory of motivation that assumes high-quality motivation is inherently human and is expressed to different degrees depending on the context that influences the process of making choices. Initial conceptualizations of motivation quality distinguished between motivations stemming from an internal locus of causality (e.g., interest and enjoyment) and those stemming from an external locus of causality (e.g., rewards and punishments) (Vansteenkiste, Lens, & Deci, 2006). A more recent conceptualization, however, distinguished among various types of extrinsic motivation that differ in their degree of autonomy, which shifted the focus to differences between autonomous motivation, which involves the experience of volition and choice, and controlled motivation, which involves the experience of being pressured or coerced (Vansteenkiste et al., 2006). Prior research has shown that learning contexts that provide choice and options for self-direction tend to facilitate autonomous motivation and enhance learning, whereas controlling environments that pressure learners to think or act in a particular way often diminish autonomous motivation and lead to poorer learning (Ryan & Deci, 2000; Vansteenkiste et al., 2004).

A common means of operationalizing autonomy-supportive and controlling learning environments is through the framing of instructions. For example, a number of laboratory and field studies have revealed that verbal or written instructions containing primarily autonomy-supportive phrases (e.g., "you may" or "if you choose") lead to higher levels of autonomous motivation and learning than instructions with more controlling phrases (e.g., "you should" or "you have to"; Vansteenkiste et al., 2004). Thus, presenting adaptive guidance instructions using autonomy-supportive language may capitalize on these motivational benefits and lead to greater learning performance than guidance instructions incorporating controlling language.

As previously noted, however, prior learner control research has revealed that greater autonomy does not always translate into higher levels of learning, and in fact sometimes leads to poorer performance (e.g., Pollock & Sullivan, 1990). A closer examination of this research suggests that these mixed findings may be due, at least in part, to differences in the learning outcomes examined across studies. For example, a meta-analysis by Patall, Cooper, and Robinson (2008) revealed small and positive effects of choice on simple task performance (i.e., quantity and accuracy), but they did not find a significant relationship between choice and subsequent measures of learning that assessed skill acquisition. Overall, they concluded that research examining the effects of choice on learning has yielded findings that have been "somewhat inconsistent" (Patall et al., 2008, p. 294). Accordingly, it may be important to consider how different forms of guidance potentially impact different types of learning outcomes. Ackerman (1987), for instance,

found that motivation and effort are the primary determinants of learners' acquisition of declarative knowledge and performance on simple tasks. Learners' motivation influences performance on basic tasks because, through practice and experience, learners develop knowledge of facts (declarative knowledge) and rules (procedural knowledge) and thus are able to perform tasks quicker and with fewer errors (Anderson, 1983). Thus, the motivational benefits of autonomy-supportive guidance should be evident on basic task components, where performance is determined primarily by effort (Bell & Kozlowski, 2002). Accordingly, we expect that learners receiving autonomy-supportive guidance will acquire basic skills at a faster rate than learners receiving controlling guidance.

Hypothesis 2: Participants in the autonomy-supportive condition will exhibit more positive change in basic performance skills than participants in the controlling condition.

The positive effects of choice in learner-controlled training, however, may not extend to learning outcomes that are a function of a trainee's ability to process and integrate complex information. Acquisition of more complex task skills is closely tied to processes related to learners' attention, such as choices made during training (e.g., sequence of study) and the quality of practice (Bell & Kozlowski, 2002; Brown, 2001), and guidance that is more controlling may increase the likelihood that trainees engage in appropriate study and practice activities. In addition, guidance design features that facilitate (rather than restrict) a learner's sense of autonomy increase the number of potential problem solutions and amount of information that needs to be processed. As the total amount of information increases, people must rely on less information to make choices, resulting in simplified problem-solving and decision-making processes and suboptimal outcomes (Chua & Iyengar, 2008; Payne, Bettman, & Johnson, 1993). For example, Iyengar and Lepper (2000) found that a greater number of options decreased people's ability to think about multiple solution combinations. By directing learners' attention to key elements of the task and limiting learners' choices, controlling guidance may enhance the acquisition and integration of skills for performing more complex components of the task. Thus, we expect that learners receiving controlling guidance will acquire strategic skills at a faster rate than learners receiving autonomy-supportive guidance.

Hypothesis 3: Participants in the controlling condition will exhibit more positive change in strategic performance skills than participants in the autonomy-supportive condition.

Interactive Effects of Guidance Design and Individual Differences

Although autonomy may yield motivational benefits during training, it is also important to consider trainees' motivation when entering a training program (i.e., pretraining motivation). Pretraining motivation describes trainees' initial attitudes and intentions to exert effort toward learning the content of a training program (Noe, 1986). Pretraining motivation is different from motivation quality constructs because pretraining motivation implies attitudes and personal action (activation) directed toward learning; motivation quality constructs address the beliefs and reasons underlying different types of motivation (Vansteenkiste, Sierens, Soenens, Luy-

ckx, & Lens, 2009). Motivated action theories have shown that attitudes and intentions provide the link between beliefs and behaviors (Heckhausen & Kuhl, 1985). Indeed, learning orientation strongly and positively predicts trainees' pretraining motivation levels (Colquitt & Simmering, 1998; Klein, Noe, & Wang, 2006), and motivation to learn has been shown, in turn, to positively relate to learning outcomes (Colquitt, LePine, & Noe, 2000).

Although pretraining motivation has been shown to be a positive predictor of training outcomes, research has also revealed that individual characteristics often interact with training design to influence learning (i.e., Aptitude \times Treatment interactions). Gully, Payne, Koles, and Whiteman (2002), for example, found that trainees higher in openness to experience had, in general, higher declarative knowledge, training performance, and self-efficacy. In addition, they found that when the training was designed to encourage exploratory behaviors consistent with this dispositional characteristic, the positive relationship was strengthened. However, when the training was designed to restrict exploration, the positive effect of openness on the training outcomes was nullified. In the current study, we propose that guidance design may play a similar role in either enhancing or constraining the positive relationship between pretraining motivation and skill acquisition. In particular, autonomy-supportive guidance should support trainees' desire to take personal action toward learning the training content, thus strengthening the relationship between pretraining motivation and learning. However, guidance that is framed as controlling should contradict trainees' positive attitudes and intentions toward the training, thus weakening the relationship between pretraining motivation and learning. Consistent with our earlier arguments, we expect the interaction between guidance design and trainees' pretraining motivation will be observed for basic skill acquisition, which is determined primarily by trainees' motivation and effort.

Hypothesis 4: Pretraining motivation will be positively related to basic performance growth for participants receiving autonomy-supportive guidance, and this relationship will be weaker for participants receiving controlling guidance.

In more complex learning environments, it is important to design training to support not only trainees' motivation but also their cognition (Bell & Kozlowski, 2008). *Cognitive ability*, which is an individual's intellectual capacity, has been shown to be a potent predictor of learning (Colquitt et al., 2000; Ree & Earles, 1991). In general, individuals with higher levels of cognitive ability have greater attentional resources to devote to learning, which means they are able to absorb and retain more information than individuals with lower cognitive ability. The challenge in learner-controlled environments is ensuring that trainees allocate their attentional resources to study and practice activities that facilitate learning. DeRouin et al. (2004) suggested that when trainees are given too much control, "they may be unable to focus the majority of their attention on the subject matter of the instructional program" (p. 154), which can cause learning to suffer. Niederhauser, Reynolds, Salmen, and Skolmoski (2000), for example, examined the effects of hypertext navigation features on learning. They found that students who made extensive use of compare-and-contrast links, which were designed to provide alternate paths to information, exhibited impaired learning, whereas students who read the text in a systematic and sequential manner performed significantly better. Niederhauser et al. (2000) suggested that the

compare-and-contrast links impeded learning because they required learners to make decisions about what to read and the order in which to read information, which likely absorbed attentional resources that could no longer be directed to integrating new knowledge. Consistent with these findings, we expect that by providing learners with a clear and unambiguous path for navigating the training, controlling guidance should enable trainees to devote more of their attentional resources to learning. This should strengthen the positive relationship between cognitive ability and performance, particularly on strategic task components that require deeper comprehension and integration of task concepts. In contrast, the relationship between cognitive ability and strategic performance should be weakened when trainees are given autonomy-supportive guidance because the greater choice options may increase the chances that attentional resources are misdirected or absorbed by instructional decisions.

Hypothesis 5: Cognitive ability will be positively related to strategic performance growth for participants receiving controlling guidance, and this effect will be weaker for participants receiving autonomy-supportive guidance.

Method

Participants

Participants were 130 undergraduate students enrolled in an introductory human resource management course at a large north-eastern university who earned course credit for participation. Fifty-nine percent of the participants were male, and most (93.1%) were between 18 and 21 years old.

Task

The task used in this study was a version of TANDEM (Dwyer, Hall, Volpe, Cannon-Bowers, & Salas, 1992), a computer-based radar-tracking simulation designed for assessing judgment and decision making in complex task environments. The object of the simulation was to make correct decisions about unknown—and potentially hostile—contacts appearing on a simulated radar screen and to prevent contacts from crossing defensive perimeters. Participants were required to detect, identify, and act on the multiple contacts on the screen using a number of basic and strategic skills (Bell & Kozlowski, 2002; Kozlowski & Bell, 2006). All participants had access to an online instruction manual that contained complete information on all important aspects of the simulation.

Basic skills involved making decisions about contacts on the radar screen. After engaging a contact, participants could access cue information from pull-down menus, with three cues available for each of three component decisions regarding the Type (air, surface, submarine), Class (civilian or military), and Intent (hostile or peaceful) of the contact. After making the three component decisions, participants needed to decide whether to take action against the contact (if hostile) or clear it from the radar screen (if peaceful). Participants received points for correct decisions and lost points for incorrect decisions.

The basic skills serve as the foundation for developing more strategic skills focused on perimeter defense and contact prioritization. Specifically, there are two defensive perimeters located within the task, and participants lose points for perimeter intru-

sions. The inner defensive perimeter is clearly marked and easy for participants to identify. However, the outer perimeter is beyond the initial viewing range of the radar display and is not clearly marked. Thus, participants must learn how to “zoom out” and locate “marker contacts” that serve to identify the outer boundary. Participants must also learn how to prioritize contacts by determining which constitute the greatest threats to the defensive perimeters. There are often multiple contacts approaching both the inner and outer perimeter, so participants need to monitor both perimeters and gather information on the speed and distance of contacts in order to determine those that are the highest priority. Trainees also have to make strategic decisions about trade-offs between contacts approaching the inner and outer perimeters, based on the number of contacts at each perimeter and their “cost” if they penetrate.

Manipulations

Learners can be given control over a number of different aspects of their instruction, including content, sequence, and pace (Kraiger & Jerden, 2007). In the current study, all trainees were given control over what they chose to study and practice (content) and the order in which they chose to study and practice the material (sequence). In addition, they were given some control over the pace of their learning, such as being able to exit the online manual early; however, for design reasons, we set maximum time limits on the study and practice periods. Thus, trainees in all conditions were given the same level of objective learner control.

At the beginning of the training session, participants in the no-guidance control condition were given a list of learning topics. They were told that the list covered all important aspects of the simulation and that they may want to focus on these topics during training, but what they chose to study and practice was at their discretion. Trainees in the no-guidance condition did not receive any guidance information.

Trainees in the guidance conditions received the list of learning topics, along with guidance information that could be used to help them evaluate their current progress and improve their deficiencies in the different aspects of the simulation. As described below, the framing of this information depended on whether trainees were assigned to the controlling or autonomy-supportive condition. The guidance information was delivered following the last screen of feedback presented after each trial. The guidance manipulations created for the current study were modeled from prior research (Bell & Kozlowski, 2002). The guidance was “adaptive” because the suggestions for study and practice were tailored to participants’ proficiency in the simulation.¹ The guidance focused on helping learners build basic skills early in training, before proceeding later

¹ The guidance was adaptive based on three levels of performance. Pilot data were used to set cutoff scores at the 50th and 85th percentiles to differentiate among low, medium, and high performance on different task components. Learners were not aware of the cutoff scores. If individuals scored below the 50th percentile, the guidance informed them that they had not yet learned how to perform the necessary skill or strategy and provided practice and study suggestions for improvement. For those scoring between the 50th and 85th percentile, the guidance informed them that they had reached a level of minimal performance, but needed to become more proficient. The guidance also provided suggestions on what they should study and practice to improve. For individuals exceeding the 85th percentile, the guidance informed them that they had mastered the skill or strategy and should focus on other areas in which they were still deficient.

in training to developing more strategic competencies that build on the fundamental skills.

The two guidance manipulations were created by framing the instructions for study and practice using language that either (a) was coercive and controlling (controlling guidance) or (b) emphasized choice and self-initiated behaviors (autonomy-supportive guidance). The specific phrases were identical to those used in a number of earlier studies that manipulated autonomy-supportive or controlling contexts through task instructions (e.g., Vansteenkiste et al., 2004). Specifically, the controlling guidance manipulation used explicitly controlling language through phrases such as “you have to,” “you must,” “you should,” and “you had better.” For example, participants might be told, “You must study the material in your manual on prioritization strategies.” The autonomy-supportive guidance manipulation used instruction phrases such as, “you can,” “you might,” “you may,” and “if you choose.” For example, participants in the autonomy-supportive guidance condition might be told, “You may want to study the material in your manual on prioritization strategies.” Other than the differences in the use of autonomy-supportive or controlling phrases, the two types of adaptive guidance were identical.

Measures

Pretraining motivation. At the beginning of the experimental session, participants’ pretraining motivation was measured using seven items developed by Noe and Schmitt (1986).² Items were modified to be consistent with our learning setting and were rated on a 5-point scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Sample items are “I am motivated to learn the skills emphasized in this training program” and “If I can’t understand something in the training program I will try harder.” Internal consistency reliability of the scale was .86.

Cognitive ability. At the beginning of the experimental session, participants provided their SAT or ACT scores. Research has shown that the SAT and ACT have a large general cognitive ability component (Frey & Detterman, 2004). In addition, the publishers of these tests report high internal consistency reliabilities for their measures (e.g., KR-20 = .96 for the ACT composite score; American College Testing Program, 1989) and self-reported SAT/ACT scores have been shown to correlate highly with actual scores. For example, Gully et al. (2002) found that self-reported SAT scores correlated .95 with actual scores. Individuals’ ACT or SAT scores were standardized using norms published by ACT and the College Board, and this standardized score was used as a measure of cognitive ability (College Board, 2011).

Basic and strategic task performance. Using measures that have been established in previous research using the TANDEM simulation (e.g., Bell & Kozlowski, 2002), data were collected during each training trial that allowed assessments of participants’ performance on basic and strategic aspects of the task. Basic task performance was calculated on the basis of the number of correct and incorrect decisions during the trials, the two fundamental components of participants’ score. Performance on these two aspects of the task is the result of knowledge of basic task components (e.g., decision-making cues and procedures). This measure is similar to task performance measures of accuracy often found in studies of choice effects on motivation (Patall et al., 2008). Strategic task performance was composed of the number of times

participants zoomed out, the number of markers hooked in an effort to identify the location of an invisible outer perimeter, and the number of high-priority contacts processed during the practice trials. These indicators capture the two major elements of strategic performance: perimeter defense and contact prioritization. Past cross-sectional research supports the two-factor structure for the performance data using TANDEM (e.g., Bell & Kozlowski, 2002).

Procedure

Training was conducted in a single 3-hr session with groups of one to four participants. During this session, participants learned to operate the radar-tracking simulation described above. Participants were randomly assigned to one of three experimental conditions: controlling guidance, autonomy-supportive guidance, or a no-guidance control condition.

Familiarization. Trainees were first presented with a brief demonstration of the simulation that described its features and decision rules and were shown the online instruction manual that contained complete information on all important aspects of the simulation. They then had an opportunity to familiarize themselves with the instruction manual for 3 min and were able to practice the task in a 5-min “familiarization” trial. The goal of this preliminary trial was to ensure that participants understood how to operate the instruction manual and were familiar with the equipment.

Training. After the familiarization trial, trainees began the training session, which was divided into nine 10.5-min trials. Each training trial consisted of a cycle of study, practice, and feedback. Trainees had 3 min to study the online instruction manual. They then had 5 min of hands-on practice. The nine trials possessed the same general profile (e.g., same difficulty level, rules, number of contacts), but the configuration of contacts (e.g., location and characteristics of contacts) was unique to each trial. Immediately after each practice trial, trainees reviewed veridical descriptive feedback on all aspects of the task relevant to both basic and strategic performance. Trainees in all conditions received feedback, but only trainees in the guidance conditions received the adaptive guidance information following the last screen of feedback in each trial. Trainees in all conditions were given the same amount of time (2.5 min) after each practice trial to review their feedback and, if available, guidance information. Participants were given a 5-min break following the third and ninth trials.

Manipulation Checks

At the end of training, all participants responded to a three-item measure of autonomous motivation adapted from Vansteenkiste et al. (2004). The items were assessed on a 5-point Likert scale that

² Pretraining motivation was assessed with eight items adapted from Noe and Schmitt (1986). Prior to modeling the latent growth trajectories, we conducted an exploratory factor analysis for the scales. One reverse-coded item, “My primary goal for this experiment is just to finish it so I get my credit,” yielded loadings less than .20 on the pretraining motivation factor. Thus, this item was dropped from the measure. The utility of reverse-coded items is frequently debated among psychometric scholars (Hinkin, 1998). In addition to internal item quality issues, dropping the item is also justified on the basis of judgmental item quality concerns, given that the measure was adapted to the context and the item may have had different meaning with the respondent population (see Stanton, Sinar, Balzar, & Smith, 2002).

ranged from 1 (*not at all true*) to 5 (*very true*). A sample item is “I practiced the task because it was very interesting.” The reliability (coefficient alpha) of the measure was .93. We ran a hierarchical regression analysis, controlling for participant’s pretraining motivation, to determine whether there were differences across the three conditions on the measure of autonomous motivation. We used one-tailed tests of significance due to the directional nature of our predictions. As expected, participants in the controlling condition ($M = 2.73$, $SD = 1.21$) reported significantly lower levels of autonomous motivation than participants in the no-guidance condition ($M = 3.28$, $SD = 1.19$), $t(129) = -2.19$, $p < .05$, and marginally significant lower levels of autonomous motivation than participants in the autonomy-supportive condition ($M = 3.01$, $SD = 1.11$), $t(129) = -1.46$, $p < .10$. Autonomous motivation did not differ significantly across the autonomy-supportive and no-guidance conditions, $t(129) = 0.92$, $p > .10$, which is consistent with the fact that participants in both conditions were told they could choose what to study and practice.

Given the subtle nature of the manipulation, we also examined the amount of time participants spent in the feedback sessions. Following each trial, participants could spend up to 2.5 min reviewing their feedback and, if available, guidance information. Participants in the no-guidance condition received only feedback, whereas participants in the controlling and autonomy-supportive conditions received both feedback and adaptive guidance information. Thus, if participants in the controlling and autonomy-supportive conditions reviewed the guidance information, we would expect them to spend more time overall in the feedback sessions. The amount of time (in seconds) participants spent reviewing the pages containing feedback and guidance (if available) information across the nine trials was automatically recorded by the computer and was subjected to regression analysis, once again using one-tailed tests of significance. The results revealed that participants in the autonomy-supportive condition ($M = 616.10$, $SD = 18.90$) spent significantly more time in the feedback sessions than participants in the control condition ($M = 418.77$, $SD = 23.26$), $t(129) = 6.58$, $p < .01$, as did participants in the controlling condition ($M = 600.33$, $SD = 21.23$), $t(129) = 5.77$, $p < .01$. Time spent in the feedback sessions did not significantly differ across the two guidance conditions, $t(129) = -0.55$, $p > .10$. Furthermore, analyses examining time spent on only the pages containing feedback information revealed that participants in autonomy-supportive condition ($M = 378.47$, $SD = 14.37$) spent significantly less time than participants in the no-guidance condition reviewing feedback ($M = 418.77$, $SD = 17.69$), $t(129) = -1.77$, $p < .05$, as did participants in the controlling condition ($M = 356.41$, $SD = 16.14$), $t(129) = -2.61$, $p < .01$. The two guidance conditions did not significantly differ in amount of time spent reviewing feedback, $t(129) = -1.02$, $p > .10$. Together, these findings show that participants in the guidance conditions spent more time in the feedback sessions, and this increase was due to the time they spent reviewing the guidance, rather than feedback, information.

Analyses

We used latent growth curve analysis (LCA) to analyze the repeated measures performance data. LCA is an extension of covariance structure analysis that invokes a confirmatory factor

analytic structure on the repeated variables measured over time, where the factor loadings for the latent growth constructs determine the shape of the growth trajectories. This approach can give identical results to other growth modeling approaches (e.g., hierarchical linear modeling) but allows greater flexibility (Curran, 2003). In particular, the latent growth curve framework allowed us to (a) test measurement invariance assumptions across time and (b) estimate growth across the three experimental conditions simultaneously by specifying a multiple-group growth curve model. Hypotheses were tested by sequentially imposing constraints on latent means (Hypotheses 1, 2, and 3) and structural paths (Hypotheses 4 and 5) and comparing nested models with the chi-square difference test (Bentler & Bonett, 1980). M-Plus was used to conduct all analyses (Muthén & Muthén, 2007). Performance measures were standardized across the nine trials. For all models, we specified autocorrelated error terms for performance scores at each time period because scores at adjacent time periods were nonindependent.

Results

Table 1 reports descriptive statistics and intercorrelations among the study variables. Inspection of the means for the basic and strategic performance outcomes shows that participants improved over time, but at a decreasing rate. Table 2 presents the basic and strategic task performance means for each condition for each of the nine training trials.

Nature of Performance Trajectories

The first step in LCA is to describe the nature of change for all participants in the sample. Table 3 presents fit statistics and nested comparisons for alternate growth trajectories (i.e., no growth, linear, and quadratic growth) and error structures (i.e., homogeneous or heterogeneous) for basic and strategic performance. The no-growth model included only a latent intercept mean and error term, whereas additional mean and error terms are included in the linear (i.e., intercept and linear terms) and quadratic (intercept, linear, and quadratic terms) models. Consistent with other longitudinal research on learning and performance during skill acquisition (e.g., Chen & Mathieu, 2008), the nested models in Table 3 show that the quadratic growth specification best fit the longitudinal data.

Table 4 presents the parameter estimates for the quadratic growth curve models. The latent factor means describe the average shape of performance growth across the nine trials for all participants. The positive linear factor means for basic ($\mu = 0.31$, $t = 9.52$, $p < .001$) and strategic ($\mu = 0.34$, $t = 11.17$, $p < .001$) performance suggest that, on average, participants scored 0.31 and 0.34 standardized points higher in each subsequent performance trial for basic and strategic performance, respectively. However, the significant negative quadratic factor means for basic ($\mu = -0.02$, $t = -4.44$, $p < .01$) and strategic ($\mu = -0.02$, $t = -5.33$, $p < .001$) performance suggest that the marginal rates of performance improvement were declining over time. Importantly, Table 4 also shows significant variation around the intercept, linear, and quadratic factors. Thus, we next specified conditional latent curve models in order to predict the individual-level variation in performance trajectories and test the study hypotheses.

Table 1
Descriptive Statistics and Correlations

Variable	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	M	SD
1. Autonomy-supportive ^a	0.41	0.49	—														
2. Controlling ^b	0.32	0.47	-.57**	—													
3. Ability (standardized)	2.81	0.85	.12	-.06	—												
4. Pretraining motivation	3.33	0.62	-.08	.06	-.04	—											
5. Performance Trial 1	-.92	0.60	.00	-.02	.30**	.06	—										
6. Performance Trial 2	-.52	0.79	-.03	.08	.32**	.05	.41**	—									
7. Performance Trial 3	-.27	0.91	-.03	.15	.21*	-.11	.27**	.69**	—								
8. Performance Trial 4	-.10	0.93	-.03	.06	.17	-.05	.22*	.62**	.73**	—							
9. Performance Trial 5	0.05	0.92	-.03	.06	.20	.03	.18*	.52**	.64**	.73**	—						
10. Performance Trial 6	0.26	0.89	-.05	.07	.22*	.02	.11	.51**	.65**	.67**	.82**	—					
11. Performance Trial 7	0.41	0.95	-.04	.00	.22*	.04	.22*	.51**	.64**	.67**	.78**	.83**	—				
12. Performance Trial 8	0.49	0.98	-.01	.04	.25**	-.06	.26**	.55**	.57**	.60**	.74**	.76**	.79**	—			
13. Performance Trial 9	0.59	0.92	.02	.05	.24**	-.03	.24**	.52**	.63**	.58**	.70**	.75**	.79**	.84**	—		

Note. Basic performance on vertical axis, strategic performance on horizontal axis and bolded.

^a 1 = Autonomy-Supportive Guidance, 0 = Controlling Guidance and No Guidance. ^b 1 = Controlling Guidance, 0 = Autonomy-Supportive Guidance and No Guidance.

* $p < .05$. ** $p < .01$.

Modeling Variation in Change

We modeled variation in participants' growth trajectories as functions of the experimental design (i.e., condition) and two time-invariant individual-difference factors (i.e., pretraining motivation and cognitive ability). Separate multiple-group growth curve models were estimated for basic and strategic performance. Cognitive ability was a single-indicator factor where we set the loading to the latent variable to the square root of the scale's reliability and set the error variance for the single indicator to one minus the reliability multiplied by the observed variance of the scale (Fornell & Larcker, 1981). We used random-item parcels to reduce the number of items on the pretraining motivation scale (Landis, Beal, & Tesluk, 2000) from seven to two items. The multiple-group models for basic $\chi^2(171, N = 130) = 238.63$, comparative fit index (CFI) = 0.941, Tucker-Lewis index (TLI) = 0.932, root-mean-square error of approximation (RMSEA) = 0.096, standardized root-mean-square residual (SRMR) = 0.095; and strategic, $\chi^2(176, N = 130) = 226.96$, CFI = 0.946, TLI = 0.939, RMSEA = 0.082, SRMR = 0.114, performance met conventional standards for fit statistics.

Table 5 presents the parameter estimates across the three experimental conditions. A visual inspection of Table 5 shows that pretraining motivation was significantly related to basic growth trajectories, whereas cognitive ability was a significant predictor of strategic growth. For both basic and strategic task performance outcomes, learners in all three conditions showed significant and positive linear performance improvements, and significant and negative quadratic performance declines (see Figure 2).

Hypothesis 1 predicted that learners receiving adaptive guidance would show greater gains in basic and strategic performance than learners receiving no guidance. Table 5 presents the means for basic and strategic performance outcomes across experimental conditions. We tested Hypothesis 1 by sequentially constraining growth factor means as equal—first across the no-guidance and autonomy-supportive guidance conditions and second across the no-guidance and controlling guidance conditions—and examining the associated change in the chi-square fit statistics between the nested models (Bentler & Bonett, 1980). Contrary to our expectations, all chi-square difference tests for linear and quadratic mean differences across basic and strategic performance models revealed nonsignificant differences between the autonomy-supportive and no-guidance conditions (all $ps > .10$).

Next, we compared the growth trajectories across participants receiving controlling guidance with those receiving no guidance. As predicted, participants receiving controlling guidance showed more positive linear growth in strategic performance ($\mu = 0.44$) than participants receiving no guidance ($\mu = 0.26$; $\Delta\chi^2 = 19.44$, $\Delta df = 1$, $p < .001$). However, results of the chi-square difference tests showed no other significant differences in performance trajectory means across the controlling guidance and no-guidance conditions for basic or strategic performance (all $ps > .10$). In sum, results showed that learners receiving controlling guidance had more positive strategic linear performance trajectories than participants receiving no guidance, yet no other differences in performance trajectories were evident. Accordingly, Hypothesis 1 received partial support.

Hypothesis 2 predicted that participants receiving autonomy-supportive guidance would exhibit greater basic performance

Table 2
Means and Standard Deviations for Performance Dimensions Across Time and Experimental Conditions

Variable	Autonomy-supportive guidance			Controlling guidance			No guidance		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Basic task performance									
Time 1	53	-0.92	0.65	42	-0.93	0.60	35	-0.90	0.54
Time 2	53	-0.55	0.87	42	-0.43	0.78	35	-0.58	0.70
Time 3	53	-0.31	0.88	42	-0.08	0.95	35	-0.44	0.89
Time 4	53	-0.07	0.97	42	-0.02	0.87	35	-0.24	0.95
Time 5	53	0.02	1.00	42	0.14	0.90	35	0.01	0.81
Time 6	53	0.21	0.95	42	0.36	0.89	35	0.23	0.80
Time 7	53	0.36	1.07	42	0.40	0.91	35	0.49	0.82
Time 8	53	0.47	1.08	42	0.54	0.98	35	0.44	0.84
Time 9	53	0.61	0.89	42	0.66	0.87	35	0.49	1.03
Strategic task performance									
Time 1	53	-1.07 _a	0.55	42	-0.85 _b	0.41	35	-0.95	0.51
Time 2	53	-0.65	0.42	42	-0.49	0.50	35	-0.68	0.46
Time 3	53	-0.57 _a	0.76	42	-0.27 _b	0.70	35	-0.43	0.61
Time 4	53	-0.23 _a	0.81	42	0.20 _b	0.94	35	-0.33 _a	0.57
Time 5	53	0.05 _a	0.85	42	0.43 _b	0.91	35	-0.17 _a	0.65
Time 6	53	0.27 _a	0.96	42	0.76 _b	0.98	35	-0.03 _a	0.77
Time 7	53	0.34 _a	1.02	42	0.96 _b	0.94	35	-0.05 _a	0.89
Time 8	53	0.50 _a	1.01	42	1.10 _b	1.02	35	0.18 _a	0.83
Time 9	53	0.49 _a	0.86	42	1.21 _b	1.03	35	0.12 _a	0.90

Note. Items are standardized. Means with different subscripts are different at $p < .05$.

growth than participants receiving controlling guidance. Figure 2 shows that, contrary to our prediction, we found that participants receiving controlling guidance exhibited marginally more positive linear basic performance trajectories ($\mu = 0.38$) than participants receiving autonomy-supportive guidance ($\mu = 0.28$; $\Delta\chi^2 = 3.51$, $\Delta df = 1$, $p < .10$). Hypothesis 2 was not supported. However, because the quadratic factor means were negative, indicating a decelerating trend, a negative relationship between a predictor and a quadratic growth factor suggests that higher levels of a predictor are associated with less deceleration in performance over time. The quadratic factor mean for participants receiving autonomy-

supportive guidance ($\mu = -0.01$) was marginally less negative than for participants receiving controlling guidance ($\mu = -0.02$; $\Delta\chi^2 = 3.18$, $\Delta df = 1$, $p < .10$), suggesting that learners' receiving autonomy-supportive guidance improved in their basic task skills at a more consistent rate than did participants receiving controlling guidance. Figure 2 shows that the basic performance differences between participants receiving autonomy-supportive guidance and controlling guidance become smaller over time.

Hypothesis 3 predicted that participants receiving controlling guidance would exhibit greater strategic performance growth than participants receiving autonomy-supportive guidance. Figure 2 (lower fig-

Table 3
Fit Statistics for Intraindividual Growth Trajectories

Model	χ^2	<i>df</i>	CFI	TLI	RMSEA	SRMR
Basic performance						
No-growth heteroscedastic	455.37***	35	0.56	0.55	0.30	0.84
No-growth homoscedastic	605.09***	43	0.41	0.51	0.32	0.41
Linear heteroscedastic	123.97***	32	0.90	0.89	0.15	0.17
Linear homoscedastic	151.96***	40	0.88	0.90	0.15	0.16
Quadratic heteroscedastic	47.27***	28	0.98	0.97	0.07	0.06
Quadratic homoscedastic	79.18***	36	0.96	0.96	0.10	0.09
Strategic performance						
No-growth heteroscedastic	926.92***	37	0.00	-0.01	0.43	2.92
No-growth homoscedastic	733.75***	43	0.20	0.33	0.35	0.76
Linear heteroscedastic	178.06***	34	0.83	0.82	0.18	0.15
Linear homoscedastic	174.99***	40	0.84	0.86	0.16	0.19
Quadratic heteroscedastic	46.25***	30	0.98	0.98	0.07	0.08
Quadratic homoscedastic	94.89***	36	0.93	0.93	0.11	0.12

Note. The degrees of freedom are different between some basic and strategic performance models. This was necessary because we found the strategic performance models with heteroscedastic error specifications arrived at improper solutions with negative uniqueness estimates for performances at Trial 9. Given the small sample size, we followed the recommendations of Gerbing and Anderson (1987) and fixed this residual to zero, which has minimal practical influence on parameter estimates or fit statistics. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual. Boldface type indicates best fitting models.

*** $p < .001$.

Table 4
Growth Curve Parameters for the Quadratic Models

Growth parameter	Basic performance		Strategic performance	
	Parameter	<i>t</i>	Parameter	<i>t</i>
Intercept				
Mean	-0.89	-16.99***	-0.95	-23.73***
Variance	0.20	2.70**	0.11	1.99*
Linear				
Mean	0.31	9.52***	0.34	11.17***
Variance	0.09	4.50***	0.09	4.86***
Quadratic				
Mean	-0.02	-4.44***	-0.02	-5.33***
Variance	0.00	3.91***	0.00	5.28***
Covariances				
Intercept with linear	-0.03	-0.87	-0.02	-0.93
Intercept with quadratic	0.00	1.00	0.00	1.14
Linear with quadratic	-0.01	-4.24***	-0.01	-4.76***

* $p < .05$. ** $p < .01$. *** $p < .001$.

ure) shows that, as expected, participants receiving controlling guidance ($\mu = 0.44$) showed greater linear growth in strategic performance than participants receiving autonomy-supportive guidance ($\mu = 0.32$; $\Delta\chi^2 = 8.79$, $\Delta df = 1$, $p < .01$). The strategic quadratic factors for controlling guidance ($M = -0.02$) and autonomy-supportive guidance ($\mu = -0.02$) were not different ($\Delta\chi^2 = 0.42$, $\Delta df = 1$, *ns*). Thus, Hypothesis 3 was supported.

Hypothesis 4 predicted that pretraining motivation will be positively related to basic performance growth for participants receiving autonomy-supportive guidance, and this relationship will be weaker for participants receiving controlling guidance. Table 5 shows that participants' pretraining motivation was positively related to basic linear growth in the autonomy-supportive guidance condition ($\beta = .27$, $EST/SE = 2.32$, $p < .05$) and negatively related to performance growth in the controlling guidance condition ($\beta = -.28$, $EST/SE = -2.21$, $p < .05$). This difference was significant ($\Delta\chi^2 = 15.84$, $\Delta df = 1$, $p < .05$) and is illustrated in Figure 3, where we plotted the interactive effects following Aiken and West's (1991) procedures. Table 5 also shows that participants' pretraining motivation was more strongly and negatively related to quadratic change (i.e., deceleration) for participants receiving autonomy-supportive guidance ($\beta = -0.03$, $EST/SE = -2.16$, $p < .05$) than for participants receiving controlling guidance ($\beta = 0.02$, $EST/SE = 1.80$, $p < .10$; $\Delta\chi^2 = 13.75$, $\Delta df = 1$, $p < .05$). This suggests that participants receiving autonomy-supportive guidance with greater pretraining motivation were able to improve their basic performance scores at a more constant rate throughout the training. Finally, as expected, Table 5 shows that participants' pretraining motivation was not significantly related to strategic performance growth in either guidance condition. These results support Hypothesis 4.

Hypothesis 5 predicted that cognitive ability will be positively related to strategic performance growth for participants receiving controlling guidance, and this effect will be weaker for participants receiving autonomy-supportive guidance. Table 5 shows that ability was positively related to linear strategic performance growth for participants receiving controlling guidance ($\beta = 0.14$, $EST/SE = 2.30$, $p < .05$) but negatively and not significantly related to performance for participants receiving autonomy-supportive guidance ($\beta = -0.05$, $EST/SE = -0.60$, *ns*). The

structural paths between ability and the linear growth factors were marginally different across the experimental conditions ($\Delta\chi^2 = 3.48$, $\Delta df = 1$, $p < .10$). To help interpret the interaction effects across guidance conditions, we plotted the interactions using Aiken and West's (1991) procedures (see Figure 4), using one standard deviation differences in participants' ability. There was also a marginally significant negative relationship between ability and the strategic performance quadratic factor for participants receiving controlling guidance ($\beta = -0.01$, $EST/SE = -1.89$, $p < .10$), suggesting that higher ability participants receiving controlling guidance were better able to sustain positive gains in strategic performance throughout the nine trials (see Figure 4). Ability was not related to the quadratic factor for participants receiving autonomy-supportive guidance ($\beta = .00$, $EST/SE = 0.16$, *ns*), and the two guidance conditions did not differ in the effect of the ability on quadratic change ($\Delta\chi^2 = 1.88$, $\Delta df = 1$, *ns*). Finally, as expected, Table 5 shows that ability was not significantly related to participants' basic performance growth in either guidance condition. Overall, these results provide support for Hypothesis 5.

Table 5
Parameter Estimates Across Experimental Conditions

Variable	AG	CG	NG
Basic performance			
Means			
Intercept	-0.88*	-0.82*	-0.87*
Linear	0.28*	0.38*	0.30*
Quadratic	-0.01*	-0.02*	-0.02*
Structural paths			
Pretraining motivation → Basic intercept	-0.01	0.07	0.02
Pretraining motivation → Basic linear	0.27*	-0.28*	0.07
Pretraining motivation → Basic quadratic	-0.03*	0.02†	-0.01
Ability → Basic intercept	0.32*	0.27*	0.13
Ability → Basic linear	-0.03	0.07	-0.04
Ability → Basic quadratic	0.00	-0.01	0.00
Strategic performance			
Means			
Intercept	-0.99*	-0.88*	-0.93*
Linear	0.32*	0.44*	0.26*
Quadratic	-0.02*	-0.02*	-0.02*
Structural paths			
Pretraining motivation → Strategic intercept	-0.22†	0.12	-0.04
Pretraining motivation → Strategic linear	0.15	-0.15	0.06
Pretraining motivation → Strategic quadratic	-0.01	0.01	-0.01
Ability → Strategic intercept	0.13	0.03	0.01
Ability → Strategic linear	-0.05	0.14*	0.04
Ability → Strategic quadratic	0.00	-0.01†	0.00

Note. Basic performance model: $\chi^2(171, N = 130) = 238.63$, CFI = 0.941, TLI = 0.932, RMSEA = 0.096, SRMR = .095; Strategic performance model: $\chi^2(176, N = 130) = 226.96$, CFI = 0.946, TLI = 0.939, RMSEA = 0.082, SRMR = .114. Modification indices suggested correlating Performance Trial 2 and 4 residuals in the NG basic model. Performance Trial 4 occurred immediately following a short break and may reasonably have impacted participants not receiving structured guidance. This change improved fit ($\Delta df = 1$, $\Delta\chi^2 = 13.71$, $p < .01$) in the basic model, but did not change any results in this study. AG = autonomy-supportive guidance; CG = controlling guidance; NG = no guidance. CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual. Predicted relationships are bolded.

† $p < .10$. * $p < .05$.

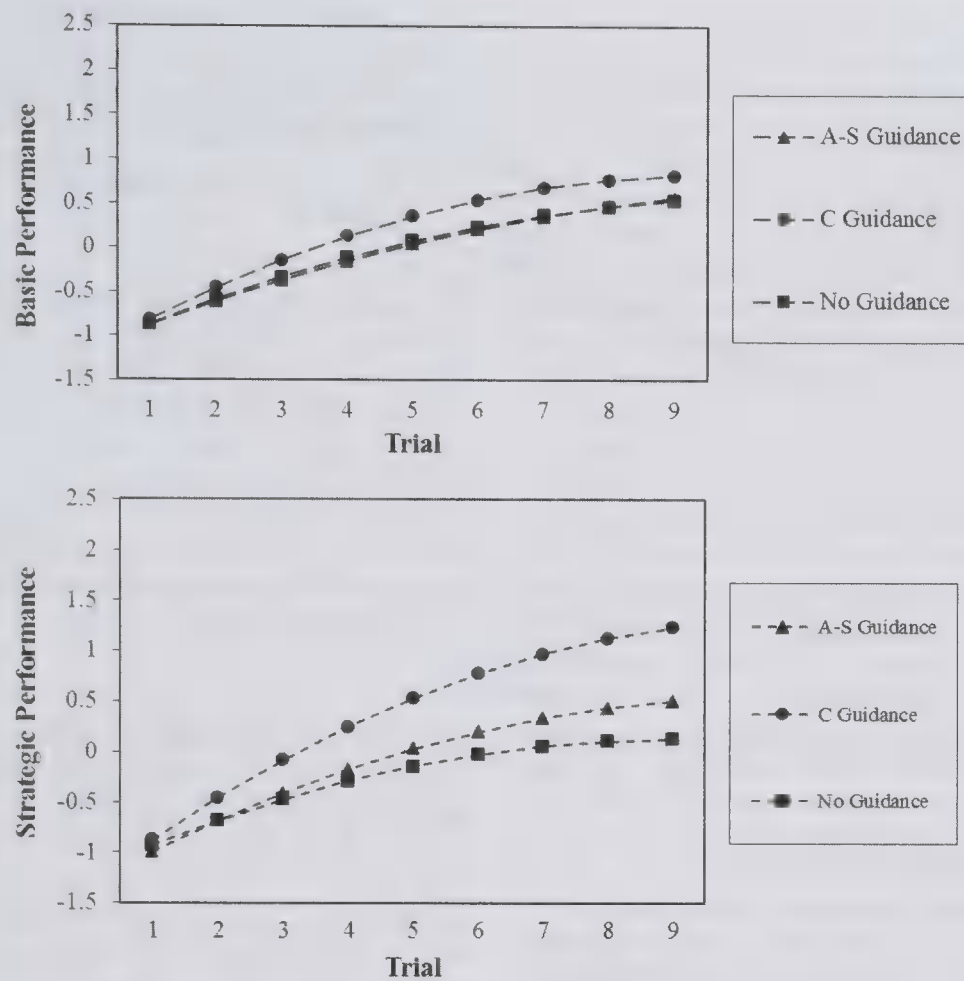


Figure 2. Mean basic and strategic performance trajectories across experimental conditions. A-S Guidance = Autonomy-Supportive Guidance; C Guidance = Controlling Guidance.

Discussion

Although educational institutions and work organizations are increasingly using computers to deliver instruction, learners often do not make good use of the control inherent in modern learning technologies (Brown, 2001). Prior research suggests that adaptive guidance can assist learners in making more

effective learning choices and can enhance learning outcomes in technology-based instruction (Bell & Kozlowski, 2002). The current investigation provides further support for the utility of adaptive guidance, but more importantly it advances research in this area by showing that the effects of guidance may vary across different design features, learning outcomes, and learner profiles. In the following sections, we review the key findings

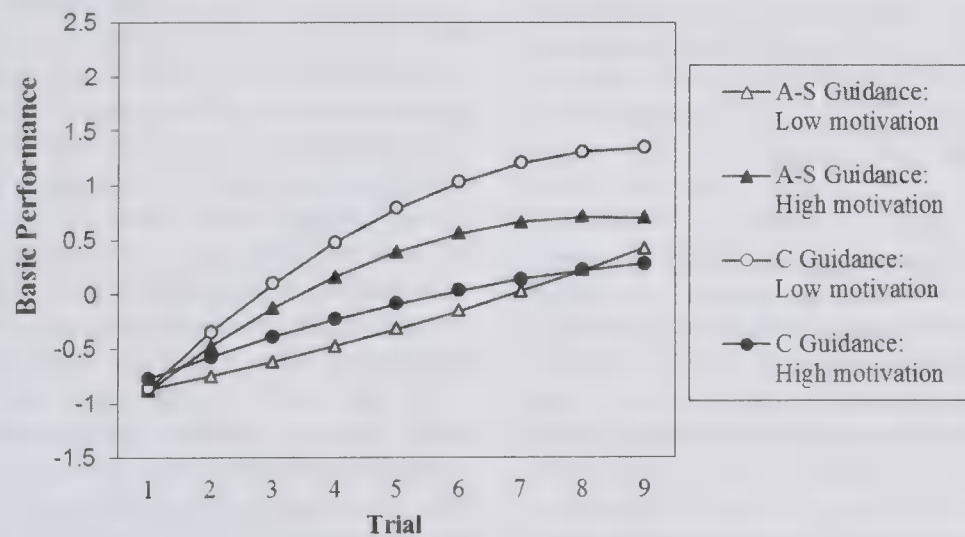


Figure 3. Influence of pretraining motivation on basic performance trajectories across experimental conditions. A-S Guidance = Autonomy-Supportive Guidance; C Guidance = Controlling Guidance.

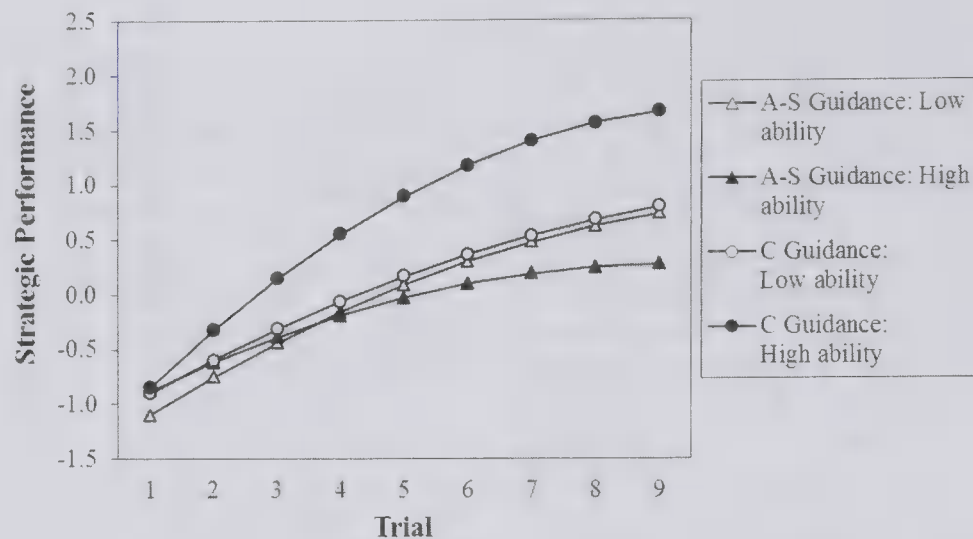


Figure 4. Influence of ability on strategic performance trajectories across experimental conditions. A-S Guidance = Autonomy-Supportive Guidance; C Guidance = Controlling Guidance.

of the current study and discuss their theoretical and practical implications.

Key Findings and Theoretical Implications

Prior research on adaptive guidance has tended to treat its effects on learning as static. To address this limitation, we used a longitudinal design and LCA to examine the effects of adaptive guidance on learning over time. The results revealed that learners who received adaptive guidance exhibited more positive change in their task performance over time than those who received no guidance, but this effect was limited to the effects of controlling guidance on strategic task performance. Adaptive guidance is designed primarily to impact the quality of learning (Bell & Kozlowski, 2002), so it is not surprising that its effects would be most pronounced for strategic performance outcomes, which are closely tied to processes related to learners' attention and require the integration of concepts and the development of task strategies. Furthermore, although we expected that both autonomy-supportive and controlling guidance would lead to more positive strategic task performance change than no guidance, the observed pattern of findings support the argument that increasing the level of direction and constraining learner choices may enhance strategic learning outcomes by reducing demands on learners' attentional resources and making it more likely that learners will come into contact with critical to-be-learned material (Kirschner et al., 2006; Mayer, 2004).

The direct comparison of autonomy-supportive and controlling guidance provided further evidence for the superiority of controlling guidance in the current context. As expected, individuals receiving controlling guidance exhibited greater linear growth in their strategic task performance than those who received autonomy-supportive guidance. Contrary to our predictions, individuals who received controlling guidance also exhibited marginally more positive basic task performance trajectories than those receiving autonomy-supportive guidance. It is important to note, however, that the basic performance trajectories of those in the controlling guidance condition showed a trend toward greater deceleration in performance growth than those in the autonomy-

supportive guidance condition (see Figure 2). Thus, future research may investigate these findings further to determine whether guidance that emphasizes autonomy and choice may lead to higher levels of basic performance when learning is extended over a longer time frame, perhaps by sustaining individuals' motivation and effort (e.g., Moller, Deci, & Ryan, 2006). Overall, however, these findings suggest that controlling guidance may be a more effective strategy for supporting skill development in more complex learning environments. Future research is needed to replicate and extend these findings, with particular attention devoted to examining the learning processes that may help further elucidate the effects of different guidance designs on various learning tasks.

A final issue examined in the current study was the interactive effects of learner characteristics and guidance design on learning over time. Drawing on SDT and resource allocation theory, we argued that individual differences related to effort (pretraining motivation) and the availability of attentional resources (cognitive ability) may interact with autonomy-supportive and controlling guidance, respectively, to influence learning trajectories. As expected, the results revealed that individuals with high levels of pretraining motivation exhibited greater growth in basic task performance when given autonomy-supportive rather than controlling guidance. Controlling guidance was detrimental to the basic task performance growth of individuals with high levels of motivation (see Figure 3), but interestingly it enhanced the performance of individuals with low levels of pretraining motivation (a finding we discuss more below). Overall, these findings suggest that autonomy-supportive guidance may support the natural expression of high-levels learning motivation, whereas controlling guidance may be effective for inducing effort from those trainees who have less positive initial attitudes and intentions toward training.

We also found that ability interacted with guidance design to impact strategic task performance. Among those who received controlling guidance, there was a positive relationship between ability and strategic performance growth. These findings support our argument that controlling guidance enables learners to allocate more of their attentional resources toward study and practice activities that will allow them to master complex task elements.

However, when individuals received autonomy-supportive guidance, ability was unrelated to strategic performance. This is consistent with our hypothesis that increasing learner choice options may absorb or divert attentional resources that could otherwise be directed toward skill acquisition.

Practical Implications

The current study suggests that the relative advantage of autonomy-supportive instructional designs relative to controlling designs may be limited in more complex tasks, and motivational guidelines alone are not sufficient for instructional design. Instead, designers should consider the extent to which the instructional program aims to teach basic or strategic skills. For basic task performance, autonomy-supportive guidance had an advantage over controlling guidance, but only for learners who possessed high levels of pretraining motivation (i.e., learners 1 *SD* above the mean; see Figure 3). This is consistent with our argument that autonomy-supportive learning contexts facilitate, and controlling contexts thwart, the beneficial effects of pretraining motivation.

Although controlling instructional designs are less frequently advocated, the current study showed a clear advantage for controlling guidance over autonomy-supportive guidance for strategic skill acquisition. Learners receiving controlling guidance showed greater gains in strategic performance than participants receiving either autonomy-supportive guidance or no guidance (see Figure 2). Furthermore, controlling guidance enhanced the positive relationship between cognitive ability and strategic performance, whereas cognitive ability was not significantly related to performance improvements for those receiving autonomy-supportive guidance. Although unexpected, the greatest growth in basic performance was observed among participants who were low in pretraining motivation who were given controlling guidance instructions. Together, these findings provide several examples of the potential utility of guidance instructions that are controlling instead of autonomy supportive.

Katz and Assor (2007) pointed out that SDT is a theory of three human needs—autonomy, competence, and relatedness. Providing choice can have implications for learning if it changes the extent to which any of these needs are or are not satisfied. Katz and Assor (2007) noted the potential resource limitations associated with providing learners with autonomy during complex tasks and suggested that instructional designers might reduce the complexity of the task to match a person's cognitive ability. On complex tasks, learners' need for competence may be more salient than their need for autonomy. In the current study, controlling guidance information may have helped to conserve attentional resources that could be directed to learning important material, thus supporting learners' need for competence. Future research can investigate whether tailoring guidance to different needs (autonomy, relatedness, and competence) can enhance the beneficial effects of adaptive guidance on learning and performance across different learning contexts. For example, Katz and Assor (2007) noted that providing choice to teams can impact relatedness needs.

Limitations and Future Research Directions

It is important to highlight a few limitations to the current research. First, the synthetic task and student sample may limit the

generalizability of our findings. Future research should extend our findings to different tasks, training spanning different lengths of time, and different instructional aids (e.g., intelligent tutors). Furthermore, future research should examine the relationships in other samples with varying levels of motivation and ability. For example, future research extending our findings in a field study using a sample varying on demographic and individual-difference factors (e.g., age) associated with different levels of motivation and ability would have important practical implications. Alternatively, researchers could attempt to manipulate attentional resources in an experimental study by varying the task demands across performance trials.

Second, Figure 3 reveals a negative relationship between motivation and basic performance for learners in the controlling guidance condition, which implies that the least motivated participants were acquiring basic skills at the fastest rate. This was an unexpected finding and suggests that controlling guidance did not thwart the positive effects of motivation on basic performance acquisition, but reversed the motivational effect (i.e., it was beneficial for unmotivated learners). We speculate that learners who lack the motivation to engage in study decisions may have defaulted to compliance, whereas learners with moderate levels of motivation may have reached a level of motivation that was sufficient to channel attentional resources away from the task. Future research is needed to first replicate and then extend this finding. Building on this finding, research may also consider other situations in which external control is preferable to intrinsic motivation to learn (cf. Pintrich, 2003).

In addition, our study design did not allow us to examine attrition from the training, which is an important practical problem for learner-control instructional designs (Sitzmann & Ely, 2010). This is an important consideration because scholars have found that controlling instructional designs may be associated with less task persistence than autonomy designs (e.g., Vansteenkiste et al., 2004). Therefore, it is important that future research include measures of attrition. Future research examining the impact of design features on attrition may also benefit by examining the type of motivation induced by the design features. For example, controlling guidance instructions may facilitate motivation that is introjected (e.g., internal control such as avoiding guilt) or external (compliance, satisfying external demands), and the difference may be important for measures of persistence. Finally, future research may want to examine instructional designs that shift the focus of guidance over time. For example, guidance designs that shift from controlling to autonomy supportive as training progresses may facilitate the acquisition of complex skills while also sustaining learners' motivation and effort over extended time frames.

Conclusion

A central issue facing learner-controlled educational technologies is that learners often make poor use of the control they are given. Thus, instructional strategies such as adaptive guidance aim to help learners to better use the control by facilitating key motivational (e.g., effort) and cognitive (e.g., learning choices) processes. This article suggests that slight changes in the design of adaptive guidance interact with individual differences in pretraining motivation and cognitive ability to impact the rate at which learners acquire basic and strategic task skills. Specifically, guid-

ance that was autonomy supportive appeared to facilitate (while controlling guidance reversed) the positive effects of pretraining motivation during basic skill acquisition. Guidance that was controlling was better for learning strategic skills, and appeared to facilitate the positive effects of cognitive ability on strategic skill acquisition. In contrast, when learners received guidance that was autonomy supportive, higher cognitive ability was not significantly related to the acquisition of strategic task skills. These findings highlight the importance of aligning the guidance design, individual differences, and skill outcome in learner-controlled environment.

References

- Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin*, 102, 3–27. doi:10.1037/0033-2909.102.1.3
- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- American College Testing Program. (1989). *Preliminary technical manual for the Enhanced ACT Assessment*. Iowa City, IA: Author.
- Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
- Bell, B. S., & Kozlowski, S. W. J. (2002). Adaptive guidance: Enhancing self-regulation, knowledge, and performance in technology-based training. *Personnel Psychology*, 55, 267–306. doi:10.1111/j.1744-6570.2002.tb00111.x
- Bell, B. S., & Kozlowski, S. W. J. (2008). Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. *Journal of Applied Psychology*, 93, 296–316. doi:10.1037/0021-9010.93.2.296
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606. doi:10.1037/0033-2909.88.3.588
- Black, A. E., & Deci, E. L. (2000). The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: A self-determination theory perspective. *Science Education*, 84, 740–756. doi:10.1002/1098-237X(200011)84:6<740::AID-SCE4>3.0.CO;2-3
- Brown, K. G. (2001). Using computers to deliver training: Which employees learn and why? *Personnel Psychology*, 54, 271–296. doi:10.1111/j.1744-6570.2001.tb00093.x
- Chen, G., & Mathieu, J. E. (2008). Goal orientation dispositions and performance trajectories: The roles of supplementary and complementary situational inducements. *Organizational Behavior and Human Decision Processes*, 106, 21–38. doi:10.1016/j.obhdp.2007.11.001
- Chua, R., & Iyengar, S. (2008). Creativity as a matter of choice: Prior experience and task instruction as boundary conditions for the positive effects of choice on creativity. *Journal of Creative Behavior*, 42, 164–180. doi:10.1002/j.2162-6057.2008.tb01293.x
- College Board. (2011). SAT-ACT concordance tables. Retrieved from <http://professionals.collegeboard.com/profdownload/act-sat-concordance-tables.pdf>
- Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years research. *Journal of Applied Psychology*, 85, 678–707. doi:10.1037/0021-9010.85.5.678
- Colquitt, J. A., & Simmering, M. S. (1998). Conscientiousness, goal orientation, and motivation to learn during the learning process: A longitudinal study. *Journal of Applied Psychology*, 83, 654–665. doi:10.1037/0021-9010.83.4.654
- Corbalan, G., Kester, L., & van Merriënboer, J. J. G. (2008). Selecting learning tasks: Effects of adaptation and shared control on learning efficiency and task involvement. *Contemporary Educational Psychology*, 33, 733–756. doi:10.1016/j.cedpsych.2008.02.003
- Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, 38, 529–569. doi:10.1207/s15327906mbr3804_5
- DeRouin, R. E., Fritzsche, B. A., & Salas, E. (2004). Optimizing e-learning: Research-based guidelines for learner-controlled training. *Human Resource Management*, 43, 147–162. doi:10.1002/hrm.20012
- Dwyer, D. J., Hall, J. K., Volpe, C., Cannon-Bowers, J. A., & Salas, E. (1992, September). A performance assessment task for examining tactical decision making under stress (Spec. Rep. No. 92–002). Orlando, FL: Naval Training Systems Center, Human Factors Division.
- Ford, J. K., & Kraiger, K. (1995). The application of cognitive constructs to the instructional systems model of training: Implications for needs assessment, design and transfer. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and Organizational Psychology* (pp. 1–48). Chichester, United Kingdom: Wiley.
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18, 39–50. doi:10.2307/3151312
- Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science*, 15, 373–378. doi:10.1111/j.0956-7976.2004.00687.x
- Gerbing, D. W., & Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika*, 52, 99–111. doi:10.1007/BF02293958
- Gully, S. M., Payne, S. C., Koles, K. L. K., & Whiteman, J. K. (2002). The impact of error training and individual differences on training outcomes: An attribute-treatment interaction perspective. *Journal of Applied Psychology*, 87, 143–155. doi:10.1037/0021-9010.87.1.143
- Heckhausen, H., & Kuhl, J. (1985). From wishes to action: The dead ends and short cuts on the long way to action. In M. Frese & J. Sabini (Eds.), *Goal-directed behavior: The concept of action in psychology* (pp. 134–160). Hillsdale, NJ: Erlbaum.
- Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in surveys. *Organizational Research Methods*, 1, 104–121. doi:10.1177/109442819800100106
- Iyengar, S. S., & Lepper, M. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79, 995–1006. doi:10.1037/0022-3514.79.6.995
- Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology*, 74, 657–690. doi:10.1037/0021-9010.74.4.657
- Katz, I., & Assor, A. (2007). When choice motivates and when it does not. *Educational Psychology Review*, 19, 429–442. doi:10.1007/s10648-006-9027-y
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41, 75–86. doi:10.1207/s15326985ep4102_1
- Klein, H. J., Noe, R. A., & Wang, C. (2006). Motivation to learn and course outcomes: The impact of delivery mode, learning goal orientation, and perceived barriers and enablers. *Personnel Psychology*, 59, 665–702. doi:10.1111/j.1744-6570.2006.00050.x
- Kozlowski, S. W. J., & Bell, B. S. (2006). Disentangling achievement orientation and goal setting: Effects on self-regulatory processes. *Journal of Applied Psychology*, 91, 900–916. doi:10.1037/0021-9010.91.4.900
- Kraiger, K., & Jerden, E. (2007). A meta-analytic investigation of learner

- control: Old findings and new directions. In S. M. Fiore & E. Salas (Eds.), *Toward a science of distributed learning* (pp. 65–90). Washington, DC: American Psychological Association. doi:10.1037/11582-004
- Landis, R. S., Beal, D. J., & Tesluk, P. E. (2000). A comparison of approaches to forming composite measures in structural equation models. *Organizational Research Methods*, 3, 186–207. doi:10.1177/109442810032003
- Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist*, 59, 14–19. doi:10.1037/0003-066X.59.1.14
- Miller, L. (2012). *State of the industry, 2012: ASTD's annual review of workplace learning and development data*. Alexandria, VA: ASTD.
- Moller, A. C., Deci, E. L., & Ryan, R. M. (2006). Choice and ego depletion: The moderating role of autonomy. *Personality and Social Psychology Bulletin*, 32, 1024–1036. doi:10.1177/0146167206288008
- Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide: Statistical analysis with latent variables* (5th ed.). Los Angeles, CA: Author.
- Niederhauser, D. S., Reynolds, R. E., Salmen, D. J., & Skolmoski, P. (2000). The influence of cognitive load on learning from hypertext. *Journal of Educational Computing Research*, 23, 237–255. doi:10.2190/81BG-RPDJ-9FA0-Q7PA
- Noe, R. A. (1986). Trainees' attributes and attitudes: Neglected influences on training effectiveness. *Academy of Management Review*, 11, 736–749.
- Noe, R. A., & Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. *Personnel Psychology*, 39, 497–523. doi:10.1111/j.1744-6570.1986.tb00950.x
- Patall, E. A., Cooper, G., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. *Psychological Bulletin*, 134, 270–300. doi:10.1037/0033-2909.134.2.270
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision-maker*. Cambridge, England: Cambridge University Press. doi:10.1017/CBO9781139173933
- Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95, 667–686. doi:10.1037/0022-0663.95.4.667
- Pollock, J., & Sullivan, H. (1990). Practice mode and learner control in computer-based instruction. *Contemporary Educational Psychology*, 15, 251–260. doi:10.1016/0361-476X(90)90022-S
- Radford, A. W. (2011). *Learning at a distance: Undergraduate enrollment in distance education courses and degree programs (NCES 2012-154)*. Alexandria, VA: National Center for Education Statistics.
- Rawsthorne, L. J., & Elliot, A. J. (1999). Achievement goals and intrinsic motivation: A meta-analytic review. *Personality and Social Psychological Review*, 3, 326–344. doi:10.1207/s15327957pspr0304_3
- Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than g. *Personnel Psychology*, 44, 321–332. doi:10.1111/j.1744-6570.1991.tb00961.x
- Reeves, T. C. (1993). Pseudoscience in computer-based instruction: The case of learner control research. *Journal of Computer-Based Instruction*, 20, 39–46.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78. doi:10.1037/0003-066X.55.1.68
- Sitzmann, T., Bell, B. S., Kraiger, K., & Kanar, A. M. (2009). A multilevel analysis of the effect of prompting self-regulation in technology-delivered instruction. *Personnel Psychology*, 62, 697–734. doi:10.1111/j.1744-6570.2009.01155.x
- Sitzmann, T., & Ely, K. (2010). Sometimes you need a reminder: The effects of prompting self-regulation on regulatory processes, learning, and attrition. *Journal of Applied Psychology*, 95, 132–144. doi:10.1037/a0018080
- Stanton, J. M., Sinar, E. F., Balzer, W., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, 55, 167–194. doi:10.1111/j.1744-6570.2002.tb00108.x
- Steinberg, E. R. (1977). Review of student control in computer-assisted instruction. *Journal of Computer-Based Instruction*, 3, 84–90.
- Steinberg, E. R. (1989). Cognition and learner control: A literature review, 1977–1988. *Journal of Computer-Based Instruction*, 16, 117–121.
- Tennyson, R. D. (1980). Instructional control strategies and content structure as design variables in concept acquisition using computer-based instruction. *Journal of Educational Psychology*, 72, 525–532. doi:10.1037/0022-0663.72.4.525
- Tennyson, R. D., & Breuer, K. (2002). Improving problem solving and creativity through use of complex-dynamic simulations. *Computers in Human Behavior*, 18, 650–668.
- Tennyson, R. D., & Buttrey, T. (1980). Advisement and management strategies as design variables in computer-assisted instruction. *Educational Communication and Technology Journal*, 28, 169–176.
- Van Buren, M. E., & Erskine, W. (2002). *State of the industry: ASTD's annual review of trends in employer-provided training in the United States*. Alexandria, VA: ASTD.
- Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational Psychologist*, 41, 19–31. doi:10.1207/s15326985ep4101_4
- Vansteenkiste, M., Sierens, E., Soenens, B., Luyckx, K., & Lens, W. (2009). Motivational profiles from a self-determination theory perspective: The quality of motivation matters. *Journal of Educational Psychology*, 101, 671–688. doi:10.1037/a0015083
- Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology*, 87, 246–260. doi:10.1037/0022-3514.87.2.246

Received December 15, 2011

Revision received June 3, 2013

Accepted June 24, 2013 ■

A Selective Meta-Analysis on the Relative Incidence of Discrete Affective States During Learning With Technology

Sidney D'Mello
University of Notre Dame

The last decade has witnessed considerable interest in the investigation of the affective dimensions of learning and in the development of advanced learning technologies that automatically detect and respond to student affect. Identifying the affective states that students experience in technology-enhanced learning contexts is a fundamental question in this area. This article provides an initial attempt to answer this question with a selective meta-analysis of 24 studies that utilized a mixture of methodologies (online self-reports, online observations, emoter-aloud protocols, cued recall) and affect judges (students themselves, untrained peers, trained judges) for fine-grained monitoring of 14 discrete affective states of 1,740 middle school, high school, college, and adult students in 5 countries. Affective states occurred over the course of interactions with a range of learning technologies, including intelligent tutoring systems, serious games, simulation environments, and simple computer interfaces. Standardized effect sizes of relative frequency, computed by comparing the proportional occurrence of each affective state to the other states in each study, were modeled with random-effects models. Engagement/flow was consistently found to be relatively frequent ($d_+ = 2.5$), and contempt, anger, disgust, sadness, anxiety, delight, fear, and surprise were consistently infrequent, with d_+ ranging from -6.5 to -0.78 . Effects for boredom ($d_+ = 0.19$), confusion ($d_+ = 0.12$), curiosity ($d_+ = -0.10$), happiness ($d_+ = -0.13$), and frustration ($d_+ = -2.5$) varied substantially across studies. Mixed-effects models indicated that the source of the affect judgments (self vs. observers) and the authenticity of the learning contexts (classroom vs. laboratory) accounted for greater heterogeneity than the use of advanced learning technologies and training time. Theoretical and applied implications of the findings are discussed.

Keywords: affect, emotion, learning, technology, meta-analysis

As most students and teachers will attest, learning is an affectively charged experience. Students experience *boredom* when the material does not appeal to them (low perceived value), when they have little or no choice over the learning task, when they cannot cope with task demands because challenges outweigh skills, and when they are understimulated when skills outweigh challenges (Csikszentmihalyi, 1975, 1990; Daschmann, Goetz, & Stupnisky, 2011; Pekrun, Goetz, Daniels, Stupnisky, & Perry, 2010). Students get *confused* when they have difficulty comprehending the material, when they encounter challenging impasses, and when they are unsure about how to proceed (D'Mello & Graesser, in press; VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003). *Frustration*

occurs when students repeatedly make mistakes, when they get stuck, or when important goals are blocked (Kapoor, Burleson, & Picard, 2007; Stein & Levine, 1991). Students even experience *despair* and *anxiety* when their efforts seem futile and when the consequence of failure is high (Zeidner, 2007).

This negative picture of the affective experiences that accompany learning has a complementary positive side. Students experience *interest* and *curiosity* when they encounter novelty and topics that interest them (Berlyne, 1978; Hidi, 2006; Silvia, 2009), *eureka* moments when insights are unveiled and major discoveries are made (Parnes, 1975), *delight* when challenges are conquered and goals are attained (D'Mello & Graesser, 2011), and perhaps even *flow-like* states of intense *engagement* when there are clear learning goals, an appropriate balance between challenges and skills, and immediate feedback on actions (Csikszentmihalyi, 1990).

These examples provide a sketch of how affective states can arise during learning activities. Systematic research focusing on the link between affect and learning has been rapidly progressing over the last decade in the interdisciplinary arena that encompasses the fields of psychology (Deci & Ryan, 2002; Dweck, 2006), education (Lepper & Woolverton, 2002; Meyer & Turner, 2006; Pekrun, 2010; Schutz & Pekrun, 2007), artificial intelligence in education (Calvo & D'Mello, 2011; Conati & Maclaren, 2009; Woolf et al., 2010), and, more recently, neuroscience (Immordino-Yang & Damasio, 2007). Progress toward uncovering links between affect and learning involves both theoretical development

This article was published Online First September 9, 2013.

This research was supported by the National Science Foundation (ITR 0325428, HCC 0834847, DRL 1235958) and by the Institute of Education Sciences and the U.S. Department of Education (DoE) through Grant R305A080594. Any opinions, findings and conclusions, or recommendations expressed in this article are my own and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education. I gratefully acknowledge Shazia Afzal, Areej Alhothali, Omar Al-zoubi, Ryan Baker, Rafael Calvo, Sazzad Hussain, James Lester, Payam Aghaei Pour, Peter Robinson, Didith Rodrigo, and Jennifer Sabourin for sharing their data.

Correspondence concerning this article should be addressed to Sidney D'Mello, 357 Fitzpatrick Hall, University of Notre Dame, Notre Dame, IN 46556. E-mail: sdmello@nd.edu

and empirical research on the affective states, the factors that give rise to these states (antecedents), and the impact of affect on the processes and products of learning (consequents).

Some of the theories that have emerged in this area emphasize the importance of (a) appraisals of control and value of the learning activity (Pekrun, 2010; Pekrun et al., 2010; Pekrun, Goetz, Titz, & Perry, 2002), (b) goal orientations (Hulleman, Durik, Schweigert, & Harackiewicz, 2008), (c) motivation and mind-set (Deci & Ryan, 2002; Dweck, 2006), (d) academic-risk taking (Clifford, 1988; Meyer & Turner, 2006), (e) interest development and maintenance (Ainley, 2008; Hidi, 2006), (f) the state of flow (Csikszentmihalyi, 1975, 1990), and (g) impasses, cognitive disequilibrium, and goal appraisals (Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005; Piaget, 1952; Stein & Levine, 1991). It is beyond the scope of this article to discuss each theory (even briefly), but it is important to emphasize that the various theories make a number of predictions on how affective states arise and influence learning outcomes. With the exception of test anxiety, which has dominated the scientific inquiry on affect during learning for the last several decades (Pekrun et al., 2010; Zeidner, 2007), there is a lot of theory and a dearth of data. Hence, several of these theories' hypotheses remain untested, leaving many fundamental questions about how affective states arise, morph, decay, and impact learning outcomes largely unanswered.

The last two decades have also witnessed an educational technology revolution in the form of advanced learning technologies (ALTs) including intelligent tutoring systems (Aleven, McLaren, Sewall, & Koedinger, 2009; Beal, Arroyo, Cohen, & Woolf, 2010; Graesser, Chipman, Haynes, & Olney, 2005), animations and simulations (Ainsworth, 2008; Mayer, 2005), and immersive educational games (Johnson & Valente, 2008; Sabourin et al., 2011). These systems all aim to positively impact learning by modeling student knowledge and engaging students in ways that far exceed the capabilities of the computer-assisted learning systems of the past (Corbett, 2001; VanLehn, 2011). The enhanced interactivity and human-like communication capabilities afforded by the ALTs are hypothesized to influence student affect in significant ways. Yet, with precious little data at hand, the affective impacts of technologically infused learning environments are not very well understood. Systematic research focused on answering basic questions, such as identifying the specific affective states that students experience while interacting with learning technologies and uncovering how these affective states influence learning, is still in its infancy.

There is also an engineering side to complement the scientific study of affect during learning. It has been suggested that one way to increase engagement and learning is to develop ALTs that can automatically detect and respond to student affect (du Boulay et al., 2010; Lepper & Woolverton, 2002; Picard, 1997). This is because it is presumed that affect is not merely incidental to learning but can also influence learning outcomes. As an example, consider boredom, a state that is negatively correlated with learning (Craig, Graesser, Sullins, & Gholson, 2004; Forbes-Riley & Litman, 2011b; Pekrun et al., 2010), presumably because bored students have trouble focusing their attention and actively persisting in the learning task. Once boredom emerges, it tends to be quite persistent (Baker, D'Mello, Rodrigo, & Graesser, 2010), which reduces the likelihood that students will reengage with the material. Baker et al. (2011) have shown that off-task behavior can

alleviate boredom, but off-task behavior itself is detrimental to learning (Baker, Corbett, Koedinger, & Wagner, 2004). Furthermore, bored students are more likely to experience frustration when they are forced to endure a learning session despite their ennui (D'Mello & Graesser, 2012); frustration is another affective state that is harmful to learning (Linnenbrink & Pintrich, 2002). Current ALTs primarily focus on the cognitive needs of the learner, so it might be the case that novel pedagogical and motivational strategies are required to inspire students to persist in learning despite the experience of negative affective states like boredom and frustration. Affect-sensitive ALTs are one way to achieve this goal.

In its most basic form, an affect-sensitive or affect-aware ALT could automatically sense when a student is bored, confused, anxious, frustrated, and so on, and intervene accordingly. Fully automated affect sensing uses predictive models that infer student affect by analyzing the context of the session and interaction profiles of the student (Baker et al., 2012; Conati & Maclaren, 2009; Sabourin, Mott, & Lester, 2011) and/or diagnostic models that sense affect from facial features, speech, postures, gestures, central and peripheral physiology, and textual responses (Chaouachi & Frasson, 2010; Pour, Hussein, AlZoubi, D'Mello, & Calvo, 2010). An affect-sensitive ALT has a number of paths to pursue once it has sensed a student's affective state, although the ideal affect-response strategies are most likely tied to aspects of the immediate situational context. Some possible interventions include doing nothing if the student is engaged and is on a positive learning trajectory; providing hints and just-in-time explanations when confusion or frustration is detected; and providing choice, encouraging breaks, or adjusting levels of challenge with respect to difficulty when boredom is detected. These and other affect-sensitive interventions have recently been implemented and compared to nonaffective interventions in ALTs, such as AutoTutor (D'Mello et al., 2010), ITSpoke (Forbes-Riley & Litman, 2011a), and Gaze Tutor (D'Mello, Olney, Williams, & Hays, 2012). Positive effects of affect sensitivity on learning gains have been documented in some contexts, but the jury is still out on the effectiveness of affect-sensitive ALTs across a range of learning environments, subject domains, and student populations.

In summary, the recent interest in exploring links between affect and learning, coupled with the emergence of affect-sensitive learning technologies, raises important questions about the role of affect during learning with technology. It is clear that understanding which affective states students are more likely to experience in technologically rich learning contexts is an important first step toward enriching understanding of affect during learning and is an essential step toward the development of systems that intelligently handle student affect. Unfortunately, available data on this most basic issue of identifying the affective states that naturally arise during learning with technology are somewhat sparse and scattered and are in need of systematic synthesis. This article addresses this issue by providing a selective meta-analysis of 24 studies that have systematically monitored student affect during interactions with both basic and advanced learning technologies. The primary goal in this analysis is to assess whether a set of discrete affective states can be consistently identified across a diverse set of laboratory and classroom studies that vary the learning task, the learning domain, the learning technology, the students who are engaging in the learning task, and the methodology used to monitor affect. It might

be the case that some affective states are consistently observed across a variety of contexts, populations, and methodologies, whereas others are more closely coupled to these factors. As such, a secondary goal in this analysis is to identify the factors that predict the variability in the incidence of the affective states.

It is important to emphasize that the present goal is not to compare affect in technology-infused versus more traditional learning contexts (e.g., classrooms or completing homework on pencil and paper). Quite different from this, the current focus is to assess the relative frequency of student affective states during learning with technology, determine the consistency of these relative frequencies across studies, and identify factors that explain variability in the relative frequencies. It should also be noted that the present focus is on discrete affective affect measurement models (e.g., boredom, anger) instead of dimensional affect measurement models (e.g., valence and arousal). The use of discrete versus dimensional models for affect representation has been an ongoing debate in the affective sciences community for over a century (Lench, Bench, & Flores, 2013; Lindquist, Siegel, Quigley, & Barrett, 2013). In our view, studies that focus on either affect representational scheme can yield important insights into affect and learning; however, discrete models are better poised to afford actionable affective response strategies, which is a major goal of affect-sensitive ALTSs. For example, an ALT that senses that a student is frustrated can respond more specifically by giving hints, displaying empathy, and so on, than an ALT that detects negative valence (i.e., dimensional affect) but is unable to determine if this is due to anger, confusion, frustration, or any other negative affective state. Of course, it is an open question if specific responses to each discrete affective state are needed, but this is currently the working hypothesis in the emerging field of affect-sensitive ALTSs.

Scope, Selection, and Description of Studies

Scope of Analysis

To appropriately contextualize this meta-analysis, note that the scientific research on affect during learning can be categorized into two separate strands of equal importance. These two research strands can be distinguished by virtue of scope, learning contexts, and methods used to track and analyze affect. The first strand focuses on a broad set of *academic emotions* (Pekrun, 2010), which include achievement emotions (e.g., frustration, anxiety), social emotions (e.g., pride, jealousy), topic emotions (e.g., empathy for a protagonist), and epistemic emotions (e.g., confusion and surprise). The dominant research methodology involves psychometrically grounded surveys that tap into a large set of variables that are hypothesized to be antecedents of affective states, such as achievement goals, situational interest, and self-concept. This line of research has yielded invaluable insights on affect and learning (see edited books by Schutz and Pekrun, 2007, and Pekrun and Linnenbrink-Garcia, in press, for examples) and has inspired the research community to probe deeper into the affective dimension of learning.

The second research strand focuses on more in-depth analyses of a smaller set of affective states that arise during learning in more restricted contexts (e.g., computer labs in schools and laboratory studies) and over shorter time spans that range from 30 to 90

minutes (see edited book by Calvo and D'Mello, 2011, for example research studies). Most, if not all, of this research focuses on learning with some form of technology, such as intelligent tutoring systems, serious games, simulation environments, and basic computer interfaces for problem solving, reading, and writing. This line of research primarily focuses on the achievement and epistemic emotions and sometimes the topic emotions. Social emotions are less relevant because most (but not all) of the research focuses on student-computer interactions rather than student-student interactions. Researchers in this group also use a more varied set of methodologies to track affect, such as online observations, emoteloud protocols, cued recall, coding of video data, and physiological and behavioral instrumentation (e.g., facial feature tracking, galvanic skin response, posture sensors).

The emphasis in this paper is on this second strand of research for a number of reasons. First, much of this research focuses on student affect during interactions with learning technologies, which is the main focus of this paper. Second, this research monitors affect at relatively fine-grained temporal resolutions ranging from seconds to minutes throughout a learning session. A fine-grained temporal resolution for affect sampling is important because coarser grained samples, such as measuring affect before and after a learning session, run the risk of overlooking the ebb and flow of dynamically changing affective states (Baker, Rodrigo, & Xolocotzin, 2007; D'Mello & Graesser, 2012). Monitoring affect at a fine-grained temporal resolution has the additional advantage of modeling the learning events that occur in close proximity with the affective states. For example, frustration after receiving negative feedback from a computer tutor is quite different from frustration due to the poor speech synthesis of an animated pedagogical agent. Third, this research is characterized by a mixed-method approach to measuring affect instead of an exclusive focus on self-reports. There is no consensus as to the most accurate method to measure affect; hence, a mixed-method approach encompassing the students themselves as well as online observers or offline video coding represents the most defensible position.

Another matter of scope pertains to how key terms such as learning activities, learning technologies, and affect are construed in this analysis. The present paper is quite inclusive in how these terms are used. *Learning activities* can range from text comprehension, problem solving, and argumentative writing to interacting with simulations, serious games, and intelligent tutoring systems (ITSs). A *learning technology* is any computer system that serves an educational purpose. It can be a complex educational game or a simple interface to support self-regulated learning. *Affective states* are also broadly construed and are taken to encompass relatively quick (seconds to a few minutes) experiences of both bona fide emotions (e.g., anger, fear) and blends of cognition and emotion (e.g., confusion, interest, and states of engagement with mild positive affect) but not longer term mood states (e.g., depression), dispositional affective traits (e.g., hostility), or motivational orientations (e.g., mastery approach tendencies). There is adequate theoretical justification to support this conceptualization of affect (Pekrun, 2010; Rosenberg, 1998; Silvia, 2009).

Search, Inclusion Criteria, and Power Analysis

Search. The studies were selected by searching (a) *International Journal of Artificial Intelligence in Education*, *Journal of*

Educational Psychology, Emotion, and Cognition & Emotion; (b) strictly peer-reviewed proceedings of the Intelligent Tutoring Systems (ITS), Artificial Intelligence in Education (AIED), and Educational Data Mining (EDM) conferences; (c) two edited books on emotions and learning (Calvo & D'Mello, 2011; Schutz & Pekrun, 2007); and (d) the Education Resources Information Center (ERIC) database and PsycINFO with queries consisting of the terms *affect*, *learning*, and *technology*. An additional search with Google Scholar was performed to obtain graduate theses and other publications not indexed by the major databases. Furthermore, some of the noted researchers who study affect and learning within the context of learning technologies were contacted for unpublished manuscripts.

It should be noted that this informal search strategy was adopted because a more formal search (searching major databases with keywords) was not yielding suitable results. This might be partially due to the infancy of the field but is also likely due to the fact that many of the researchers in this area tend to present their work in strictly reviewed conferences that include published proceedings in the form of edited volumes, which are not always indexed in the major databases. There is some confidence that the present informal search strategy, which included targeting specific outlets, uncovered most of the relevant articles, because the formal approach of searching ERIC and PsycINFO rarely uncovered an article that was not discovered during the targeted search.

Inclusion criteria. Twenty-four studies were selected for the analysis on the basis of the following criteria related to the learning context, method used to monitor affect, affect measurement model, sample size, and availability of data.

With respect to the learning context, the only requirement was that the studies should involve interactions with some form of learning technology. A broad definition of learning technology was adopted, as discussed above.

The following criteria pertained to the methodology used to monitor student affect. There was the requirement that student affective states should be tracked during the learning session. Studies that monitored affect only before and after a learning session were excluded. There was also the requirement that the affective states should be measured at a relatively fine-grained temporal resolution. Studies that tracked affect two or three times during a learning session were excluded because this does not afford sufficiently fine-grained monitoring of affect, which is the focus of this analysis. Additionally, studies that inferred affective states entirely on the basis of physiology, facial activity, EEG, and other signals such as posture, gesture, eye gaze, and acoustic features (e.g., Chaouachi & Frasson, 2010; Harley, Bouchet, & Azevedo, 2012) were also excluded due to open issues pertaining to the validity of fully automated affect measurement systems (Calvo & D'Mello, 2010; Kappas, 2010; Zeng, Pantic, Roisman, & Huang, 2009). Finally, studies that tracked a single student across multiple learning sessions were not included because between-student error estimates, which are required for the statistical analysis, are not available in these single-student studies.

The affect measurement model pertains to the list of affective states measured as well as the instruments used to measure these states. The present focus was on discrete affective states for the reasons noted in the introduction. A majority of the available studies focused on discrete affect, so this inclusion criterion did not drastically reduce the pool of available studies. In line with this

requirement, only studies that recorded the presence or absence of discrete affective states (e.g., boredom, anger, frustration) at multiple instances in the learning session were included. Studies that employed dimensional affect models, such as valence and arousal (e.g., Vuorela & Nummenmaa, 2004), and studies that grouped affective states into broad categories, such as positive, neutral, and negative (Litman & Forbes-Riley, 2006), were excluded. Studies that measured affect via Likert-type scales instead of binary self-reports (e.g., Arroyo et al., 2009; Conati & Maclaren, 2009) were also excluded because this introduced complications with respect to computing the effect size statistic (discussed in the Analysis and Results section). Similarly, studies that focused on the presence versus absence of a single affective state (e.g., Forbes-Riley & Litman, 2009) or on multiple levels of a single affective state, such as multiple levels of engagement but not any other state (e.g., Mota & Picard, 2003), were not included because this does not permit comparisons with other affective states.

Studies were also selected on the basis of data availability. The following three units of information were required by the statistical approach adopted for the analysis: (a) number of students, (b) mean proportional occurrence of each discrete affective state across students, (c) standard deviation of proportional occurrence of each discrete affective state across students, and (d) correlation matrix of proportional occurrence of affective states. Studies were excluded if they did not report this information in published reports or if the author(s) of the candidate studies did not respond to requests to provide the necessary information.

It is important to point out that the present analysis was intended to answer two very specific affect-related questions (see the introduction) in a very specific learning context (i.e., learning with technology). Therefore, very specific publication outlets were targeted and somewhat stringent inclusion criteria were adopted. The relatively informal nature of the search, which makes replication difficult, and the fact that only one author performed the search and applied the inclusion/exclusion criteria, does not meet recommended criteria for performing a comprehensive meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hammerstrøm, Wade, & Jørgensen, 2010). Hence, as reflected in the title, the present paper should be considered to be a "selective meta-analysis."

Power analysis. A power analysis was conducted to determine if the sample of 24 studies yielded adequate power to detect significant effects in relative affect incidence (relative because each affective state was compared to other affective states). A power analysis that used Cohen's (1992) guidelines for small ($d = 0.2$ sigma), medium ($d = 0.5$), and large effects ($d = 0.8$), Hedges and Pigott's (2001) conservative heterogeneity estimates, Borenstein et al.'s (2009) formulas for power analysis of main effects—assuming that each study had an N of 30—indicated that 22 studies would have adequate power (power = 0.8) to detect small ($d = 0.2$) or larger effects with a two-tailed test with an alpha level of 0.05. This power analysis is somewhat conservative, because the estimated N of 30 is substantially lower than the mean number of students in the studies (mean N was 73, as discussed below). Repeating the power analysis with an N of 73 indicated that 11 studies were needed to detect small or larger effects and the 24 studies could detect effects as small as 0.14 sigma.

Description of Studies

Table 1 provides an overview of the methodologies of the 24 studies that were analyzed. The subsequent discussion focuses on some of the key between-study differences.

Student populations. Student samples exhibited diversity in terms of education level, age, ethnicity, country, and native language. The samples consisted of middle school, high school, college, and adult students from the United States, Canada, the United Kingdom, Philippines, and Australia. The number of students per study ranged from 7 to 260, with a median of 39, a mean of 73, and a standard deviation of 66. In all, affect data from 1,740 students were analyzed.

Training time. Training time refers to the amount of time students engaged with the learning technology. Training times ranged from 10 min to 90 min, with a median of 39 min, and a mean of 43 min ($SD = 21$). Data from most studies were collected in a single session, but some studies had multiple sessions.

Learning contexts. The term *learning context* encapsulates the learning setting, learning task, learning technology, and learning topic. In terms of the learning setting, there was a mix of laboratory, school, and online studies. The school studies were usually conducted in computer laboratories rather than classrooms because technology was involved. Data from the three online studies with adults were collected with Amazon Mechanical Turk, which is a crowd sourcing platform that allows individuals to receive monetary compensation for completing human intelligence tasks online (<http://www.mturk.com>). Studies in which the primary learning activity was linked to the students' classroom curricula and studies conducted in computer labs in schools are grouped in the "Authentic" versus "Laboratory" context category in Table 1.

There was considerable diversity in terms of tasks, topics, and technologies. Learning tasks involved one-on-one tutoring sessions with ITSs, solving logic puzzles, interacting with simulations and serious games, developing reading comprehension, practicing for standardized tests, and developing writing proficiency. Topics (subject domains) covered included algebra, analytical reasoning, argumentative writing, chemistry, computer literacy (hardware, operating systems, the Internet), ecology, genetics, geography, graphing, logic puzzles, microbiology, pre-algebra, and social studies. The learning technologies included intelligent tutoring systems, serious games, simulation environments, virtual labs, and computer interfaces for problem solving, reading comprehension, and essay writing. Students individually worked with the learning technology in almost all of the cases (see Baker et al., 2011, for an exception).

Methodologies to monitor affective states. Affective states were tracked with a number of methodologies including online self-reports, emote-aloud protocols, online observations, and retrospective coding of video after a learning session by the students themselves or by peers, observers, or trained judges. Online self-reports involved periodically polling the student for an affect report, whereas emote-aloud protocols asked students making spontaneous (i.e., nonprompted) verbal reports on their affective states as these were consciously experienced (Craig, D'Mello, Witherspoon, & Graesser, 2008). Online observations involved one or more coders making observations on student affect during a learning session (Rodrigo & Baker, 2011a). Cued-recall or retrospective affect judgment protocols involved collecting video

recordings of a student's face and computer screen (to capture context) during the session and obtaining affect judgments over the course of replaying these videos after the session (Graesser, Chipman, King, McDaniell, & D'Mello, 2007). The affect judgments for studies using these offline retrospective judgments were usually provided by the students themselves, but other studies used observers, such as trained judges, peers, or teachers (Graesser et al., 2006).

Affect sampling rates varied as a function of methodology and study. Several studies used a fixed sampling rate, where affect measurements were collected in regular intervals ranging from 15 s to 7 min. Another option was voluntary measurements, in which affect reports were made as they occurred online or on the basis of offline video coding. Some studies even used a combination of fixed and voluntary measurement, where affect measurements were collected at fixed intervals, but judges (or raters) were permitted to offer voluntary judgments between two fixed sampling points (e.g., Graesser et al., 2006). Some studies used an event-based sampling method, in which affect measurements were elicited to correspond to predetermined events (e.g., at the start of a new problem or topic, after completion of a problem or topic, or a few seconds after receiving feedback; e.g., Graesser et al., 2007). Table 1 lists sampling rates when available, which was usually for studies that used fixed sampling methods or fixed + voluntary sampling methods. It is impossible to compute precise sampling rates for event-based and other methods, so these have simply been listed as "Varied" in the table.

Affective states considered in the studies. A total of 17 affective states plus neutral were tracked in the 24 studies (see Table 2 for an alignment of affect by study). These include anger, anxiety, boredom, confusion, contempt, curiosity, delight, disgust, engagement/flow, eureka, excitement, fear, frustration, happiness, interest, sadness, and surprise. Twenty studies included a neutral category, four studies incorporated an "other" category, and one study included a "none" category. The number of affect states included in each study (including neutral, other, and none) ranged from 5 to 15 with a mean of 8 states ($SD = 3$) and a median of 7 states.

The definitions of most of these affective states are well known, but engagement/flow requires some clarification. Quite different from passively attending to a task, engagement/flow is conceptualized as a state of mild positive affect when involved with a task, such that concentration is intense, attention is focused, and focus is complete. However, it may not involve some of the aspects of Csikszentmihalyi's (1990) conceptualization of flow that refer to extreme intensity to the extent that there is time distortion or loss of self-consciousness.

It should also be noted that fear and anxiety are related but distinct constructs, as argued by Öhman (2008).

Analysis and Results

Encoding and Computing Effect Size Statistics

There was a difference in the set of affective states monitored in each study, and some states were included in only a handful of studies. A power analysis with power = 0.8, $\alpha = 0.05$, and N of 73 (this is the mean N across the 24 studies) indicated that a minimum of 4 studies was needed to detect at least a small- to

Table 1
Synopsis of Studies Included in Selective Meta-Analysis

No.	Learning				Student population			Affect measurement				
	Task	Technology	Domain	Setting ^a	Min ^b	No. S ^c	Level	Country	No. A ^d	Method	Rate ^e	Reference
1	Interaction with ITS	AutoTutor	Computer literacy	Laboratory	45	27	College	Philippines	9	Retrospective (self)	20 s	Pour et al. (2010)
2	Writing essays	Computer interface	Creative/argumentative writing	Laboratory	30	36	College	USA	15	Retrospective (self)	15 s	Mills & D'Mello (2012)
3	Interaction with ITS	AutoTutor	Computer literacy	Laboratory	35	34	College	USA	6	Observational	300 s	Craig et al. (2004)
4	Interaction with serious game	Math Blaster	Prealgebra	Authentic	40	30	Middle	Philippines	7	Observational	200 s	Rodrigo & Baker (2011b)
5	Interaction with ITS	AutoTutor	Computer literacy	Laboratory	32	28	College	USA	7	Retrospective (self, peers, and trained judges)	20 s	Graesser et al. (2006)
6	Interaction with simulation environment	Ecolab	Ecology	Authentic	40	180	Middle	Philippines	7	Observational	200 s	Rodrigo & Baker (2011a)
7	Standardized test practice	Computer interface	Analytical reasoning	Laboratory	35	41	College	USA	14	Retrospective (self)	Varied	Lehman et al. (2008)
8	Interaction with ITS	Scatterplot tutor	Graphing	Authentic	80	126	Middle	Philippines	8	Observational	200 s	Rodrigo & Baker (2011a)
9	Map-based geography tutorial and card-sorting task	Computer interfaces	Geography and logic puzzles	Laboratory	30	8	College	U.K.	8	Retrospective (judges)	Varied	Afzal & Robinson (2011)
10	Interaction with ITS	Operation ARIES!	Research methods	Laboratory	60	64	College	USA	9	Retrospective (self)	Varied	D'Mello et al. (2012)
11	Interaction with ITS	Aplusix	Algebra	Authentic	45	140	Middle	Philippines	7	Observational	200 s	Rodrigo & Baker (2011b)
12	Interaction with ITS	AutoTutor	Computer literacy	Laboratory	35	30	College	USA	7	Retrospective (self)	Varied	Graesser et al. (2007)
13	Reading comprehension	Web interface	Social studies	Laboratory	35	131	Adults	USA	6	Online self-report	Varied	Strain & D'Mello (2011)
14	Interaction with ITS	AutoTutor	Computer literacy	Laboratory	20	20	College	Australia	8	Retrospective (self)	10 s	Sazzad et al. (2011)
15	Interaction with ITS	CompTutor	Computer-related topics	Laboratory	45	19	College	Canada	5	Retrospective (self and trained judges)	Varied	Alhothali (2011)
16	Writing essays	Computer interface	Creative/argumentative writing	Laboratory	30	44	College	USA	15	Retrospective (self)	15 s	D'Mello & Mills (in press)
17	Writing essays	Web interface	Creative/argumentative writing	Laboratory	12	166	Adults	USA	12	Online self-report	120 s	D'Mello & Mills (in press)

(table continues)

(table continues)

Table 1 (continued)

No.	Learning			Student population			Affect measurement					
	Task	Technology	Domain	Setting ^a	Min ^b	No. S ^c	Level	Country	No. A ^d	Method	Rate ^e	Reference
18	Narrative-centered learning environment	Crystal Island	Microbiology and genetics	Authentic	55	260	Middle	USA	7	Online self-report	420 s	Sabourin et al. (2011)
19	Interaction with ITS	AutoTutor	Computer literacy	Laboratory	90	7	College	USA	8	Emote-aloud	Varied	Craig et al. (2008)
20	Interaction with serious game	Incredible Machine	Logic puzzles	Authentic	10	36	High	Philippines	7	Observational	60 s	Rodrigo & Baker (2011a)
21	Interaction with ITS	Operation ARIES!	Research methods	Laboratory	60	31	College	USA	9	Retrospective (self)	Varied	Lehman et al. (2011)
22	Reading comprehension	Web interface	Social studies	Laboratory	25	138	Adults	USA	6	Online self-report	Varied	Strain & D'Mello (2011)
23	Virtual laboratory	ChemCollective: Virtual labs	Chemistry	Authentic	45	55	College	USA	4	Observational	131 s	Baker et al. (2011)
24	Interaction with ITS	Cognitive Tutor	Algebra	Authentic	90	89	High	USA	4	Observational	332 s	Baker et al. (2012)

Note. ITS = intelligent tutoring system.

^a Denotes whether affect was tracked in an *authentic* setting, where students were monitored in classrooms (most studies), or in a *laboratory* setting, where they were monitored in a lab (including online studies). ^b Min = length of learning session in minutes. ^c No. S = number of students. ^d No. A = number of affective states. ^e Affect sampling rate when applicable in seconds, with Varied indicating no fixed rate.

medium-sized effect ($d = 0.35$) under conditions of severe heterogeneity. As such, states that were present in fewer than four studies were excluded from the subsequent analyses. These included excitement, interest, and eureka.

The critical dependent variable for an affective state was the proportional occurrence of that state for a given student in a given study. Hence, the sum of proportions for a single student would add up to 1. Some studies utilized multiple judges to assess students' affective states (Alhothali, 2011; D'Mello & Graesser, 2011). This can violate independence assumptions, so proportional scores were averaged across the multiple judges prior to computing the effect sizes.

Two sets of standardized effect size measures were computed for each affect state (target affective state) from these proportion scores. The first was an *overall effect size*, which was the standardized mean difference (Cohen's d) between the proportion of the target affective state compared to the *average proportion* of the other states. This would yield one standardized overall effect size for each affective state in a study. There would be 350 effect sizes (14 states) if every study used the same affect labels. However, this was never the case, and when summed across studies, there were 159 overall effect sizes.

The second set of effect sizes, called *pairwise effect sizes*, was the standardized mean difference (Cohen's ds) when the proportion of each affective state was compared to the proportions of each other state in the study. Hence, a study with e affective states would yield $[e \times (e - 1)/2]$ pairwise effect sizes. Overall, there were 596 pairwise effects across the 24 studies.

It is important to note that the two effect size measures provide similar information on the relative incidence of each affect state, albeit at different levels of granularity. The overall effect size provides a coarse-grained assessment of the relative incidence of a target affect state by comparing it to the average of the other affective states. The pairwise effect size provides a finer grained assessment because the target affective state is compared to each other affective state in the study. Both measures are not affected by methodological differences in terms of affect measurement models across studies, because proportional scores were compared only within an individual study. That is, the sum of the proportions of the affective states in a study (including neutral, none, and other) always summed up to 1, so within-study differences, such as including a different list of affective states, would not affect the results. Additionally, a given affective state could occur with the same proportion in two studies but could have different effect sizes (both overall and pairwise), based on the other states included in the study. This is exactly what is needed because the metric of interest is "relative frequency," which should vary as the affective states considered in each study vary, as opposed to an absolute measure, which should be the same across studies. Furthermore, a standardized (instead of a raw) effect size estimate was used in the analyses, which is widely recommended in the literature because it makes it feasible to aggregate effects across studies (Borenstein et al., 2009; Lipsey & Wilson, 2001). That being said, the analyses were repeated with an unstandardized effect size metric (the raw proportions for each state) and the same major patterns were discovered.

A fully automated approach was adopted in order to minimize errors associated with computing and analyzing the effect sizes. First, the authors of the 24 studies were contacted with a request to

Table 2
Alignment of Affective States by Individual Study

Affect	Study no.																								N studies
	1 ^a	2	3	4	5 ^a	6	7	8	9	10	11	12	13	14	15	16	17 ^a	18	19	20	21	22	23	24	
Anger		+					+		+							+	+		+						6
Anxiety		+					+			+						+	+	+			+				7
Boredom	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	24
Confusion	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	24
Contempt		+					+									+			+						4
Curiosity	+	+					+			+				+		+		+	+		+				9
Delight	+	+		+	+	+		+		+	+	+		+	+	+				+	+				14
Disgust		+					+									+	+		+						5
Engagement/flow	+	+	+	+	+	+		+		+	+	+	+	+		+	+	+		+	+	+	+	+	20
Eureka			+				+												+						3
Excitement																		+							1
Fear		+					+									+	+								4
Frustration	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	23
Happiness		+					+		+				+			+	+					+			7
Interest									+								+								1
Sadness		+					+									+	+								4
Surprise	+	+		+	+	+	+	+	+	+	+	+		+		+	+			+	+				16
Neutral	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			+	+	+			20
Other	+								+														+	+	4
None								+																	1
N states	9	15	6	7	7	7	14	8	8	9	7	7	6	8	5	15	12	7	8	7	9	6	5	5	

Note. Please see Table 1 for details on individual studies indexed by study number.

^a Study methodology permitted multiple affective states to be simultaneously reported.

provide student-level proportionalized affective state data. This consisted of an $n \times e$ (*student* \times *affective state*) matrix with cell (i, j) representing the proportional occurrence of affective state j for student i . A computer program was developed to compute the within-subject standardized effect size and its variance with formulas specified in Borenstein et al. (2009). After accuracy was verified, the program was used to automatically compute the 159 overall and 596 pairwise effect sizes along with their variances.

An examination of the effect size distributions revealed a few outliers that would potentially skew the results. Outliers were identified as d s that exceeded three standard deviations from the mean and were removed as recommended by Lipsey and Wilson (2001). A total of 3 overall effects (1.89%) and 26 (4.5%) pairwise effects were identified as outliers and removed.

The metafor package (Viechtbauer, 2010), which is a validated library for conducting meta-analyses within the R environment, was used for all subsequent analyses. Restricted maximum likelihood estimation (REML) was used for model fitting. Both overall and pairwise effects were analyzed and are reported here; however, the emphasis is on overall effects, because they provide a clearer picture on the relative incidence of the affective states.

Random-Effects Models on Overall Effect Sizes

A random-effects model was used to model the effect size distributions due to the considerable between-study variability in student populations, learning technologies, and methodologies. A random-effects model assumes that the true effect varies across studies due to between-study variance (τ^2) and within-study variability. Total variance for each study is the sum of between- (constant for all studies) and within-subject variability.

The homogeneity statistic (Q ; Cochran, 1954) was used to test whether the true effect sizes were consistent (homogeneity) or varied across studies (heterogeneity), with significant Q s being indicative of heterogeneity. With the exception of that for contempt, Q s for all effects were significant ($p < .05$ unless with two-tailed tests specified otherwise), thereby indicating heterogeneity in effect size distributions (see Table 3). The table also lists the I^2 statistic, which is the percent of variance that can be attributed to true heterogeneity (Borenstein et al., 2009). I^2 for the heterogeneous effect sizes (i.e., those with nonsignificant Q s) ranged from 63.2% to 99.7%, which indicates that much of the variance was caused by real between-study differences instead of random error. These between-study differences are explored with moderation analyses in the next section.

Weights were assigned to each study based on inverse variance weighting, and the key summary measure was the weighted mean effect size (d_+). For affective state e , d_+ indicates the weighted mean (across studies) of the effect sizes of e , where each effect size (d) is the standardized mean difference (Cohen's d) between the proportion of e and the average proportion of the other states in each study. Descriptive statistics (in descending order of d_+) along with significance tests of the main effect (d_+) for the 14 affective states are presented in Table 3. The results supported a four-category grouping of affective states in terms of the magnitude and significance of the weighted mean effect sizes. Engagement/flow was the only state that yielded a significant and positive d_+ of 2.46. Weighted mean effects for boredom ($d_+ = 0.19$) and confusion ($d_+ = 0.12$) were consistent with small nonsignificant positive effects. Curiosity ($d_+ = -0.10$) and happiness ($d_+ = -0.13$), on the other hand, yielded small nonsignificant negative effects. The

Table 3
Descriptives of Effect Sizes and Heterogeneity Analysis Based on Random Effects Model

Affective state	k^a	Descriptive statistics of d_+				Range of d			Heterogeneity		
		$M (SE)$	95% CI [LL, UL]	Z	p	$d < -0.2$	$-0.2 \leq d \leq 0.2$	$d > 0.2$	Q^b	Tau^2	I^2
Significant and positive Engagement/flow	19	2.5 (0.46)	[1.6, 3.4]	5.37	<.01	0.00	0.00	1.0	213.4	3.6	97.4
Not significant and positive Boredom	21	0.19 (0.27)	[-0.34, 0.71]	0.69	.49	0.33	0.05	0.62	259.4	1.3	95.4
Confusion	20	0.12 (0.28)	[-0.43, 0.66]	0.42	.67	0.40	0.15	0.45	228.1	1.4	96.3
Not significant and negative Curiosity	9	-0.10 (0.32)	[-0.73, 0.52]	-0.33	.74	0.44	0.22	0.33	79.1	0.7	91.2
Happiness	7	-0.13 (0.22)	[-0.55, 0.30]	-0.60	.55	0.43	0.29	0.29	36.0	0.3	86.4
Significant and negative Contempt	4	-0.78 (0.14)	[-1.1, -0.50]	-5.47	<.01	1.0	0.00	0.00	2.5	0	0
Anger	6	-1.2 (0.34)	[-1.9, -0.55]	-3.59	<.01	1.0	0.00	0.00	28.7	0.4	80.6
Disgust	5	-1.5 (0.44)	[-2.3, -0.61]	-3.35	<.01	1.0	0.00	0.00	17.0	0.7	89.2
Sadness	4	-1.5 (0.25)	[-2.0, -1.0]	-6.13	<.01	1.0	0.00	0.00	9.1	0.1	63.2
Anxiety	7	-1.5 (0.68)	[-2.9, -0.18]	-2.23	.03	0.71	0.29	0.00	138.5	3.1	98.1
Delight	12	-2.1 (0.37)	[-2.8, -1.4]	-5.64	<.01	1.0	0.00	0.00	88.8	1.5	90.8
Frustration	21	-2.5 (1.1)	[-4.6, -0.38]	-2.31	.02	0.52	0.19	0.29	378.9	23.7	99.7
Fear	4	-2.7 (0.64)	[-4.0, -1.5]	-4.29	<.01	1.0	0.00	0.00	21.2	1.4	89.7
Surprise	14	-6.5 (2.0)	[-10.3, -2.6]	-3.30	<.01	0.93	0.00	0.07	330.8	52.3	99.4

Note. SE = standard error; CI = confidence interval; LL = lower limit; UL = upper limit.

^a k is the number of studies. ^b All Q s significant at $p < .001$ except for contempt, where $p = .048$.

remaining nine affective states (anger, disgust, sadness, anxiety, contempt, delight, frustration, fear, and surprise) had significantly negative d_+ s ranging from -6.5 to -0.78.

Although d_+ provides a useful summary measure of the overall relative incidence of each affective state across all the studies, it is somewhat insensitive to the nuances in the individual effect size distributions. This was investigated by categorizing the individual effects on the basis of Cohen's (1992) proposed convention of 0.2, 0.5, and 0.8 sigma representing small, medium, and large effects, respectively. In particular, each effect was grouped into one of the following three categories: (a) small or larger negative effect ($d < -0.2$), (b) negligible effect ($-0.2 \leq d \leq 0.2$), and (c) small or larger positive effect ($d > 0.2$). The proportion of studies falling into each of the three categories is presented in Table 3. This analysis indicated that the effect sizes of engagement/flow were always greater than 0.2, and the effect sizes of contempt, anger, disgust, sadness, delight, and fear were consistently less than -0.2. With minor exceptions, the effect sizes for surprise and anxiety were less than -0.2. The results for these nine states were consistent with the patterns of the weighted mean effect sizes.

The data were more interesting for boredom, confusion, curiosity, happiness, and frustration. Boredom and confusion were two affective states with positive nonsignificant d_+ s, yet d s for these states alternated between small or larger negative ($d < 0.2$) and positive ($d > 0.2$) effects. Effect sizes for curiosity and happiness, which were two states with nonsignificant negative d_+ s, alternated among all three categories ($d < -0.2$; $-0.2 \leq d \leq 0.2$; $d > 0.2$). The data for frustration were somewhat surprising, as the d_+ s for this state was significant and negative, but d s > 0.2 were observed for approximately one third of the studies. Taken together, it

appears that the low d_+ s for boredom, confusion, curiosity, happiness, and frustration should not be attributed to consistently low d s across studies but rather to considerable between-study variability in d s. This is different from the other nine states that consistently yielded positive effects (flow/engagement) or negative effects (contempt, anger, disgust, sadness, delight, fear, anxiety, and surprise).

Mixed-Effects Models on Overall Effect Sizes

Mixed-effects models assume that a portion of between-study variability can be modeled by considering systematic differences between studies (i.e., moderators). The four study-level moderators that were considered were *self-report*, *authentic context*, *ALT used*, and *training time* (see Table 1 for details on these variables). Self-report was a dichotomous variable that indicated whether the affect measurements were provided by the students themselves (coded as 1) or by an observer (coded as 0), such as a trained judge, a researcher, or an untrained peer. Authentic context was also a dichotomous variable; it was coded as 1 if the learning context occurred in a school setting and as 0 if the context occurred in laboratory studies. ALT was dichotomous and represented whether the learning environment was an advanced learning technology, such as an ITS or a serious game (coded as 1), or a simpler computer interface (coded as 0). Finally, training time was a continuous variable that represented the length of the training session in minutes.

Additional study-level variables from Table 1 (i.e., learning task, learning technology, task domain, population, and country of students) were not considered because these categorical variables had too many levels, thereby making it difficult to derive mean-

ingful models given the constraints of the data. Student population was not included as an independent variable because it was highly correlated with *authentic context*. That is, college students were the primary participants in laboratory studies, but middle and high school students made up the samples of studies in more authentic learning contexts (i.e., in schools).

Tolerance values for the four moderator variables ranged from .361 to .451. This suggests that there were no severe multicollinearity problems, because tolerances exceeded or were very close to the recommended value of 0.4 (Allison, 1999).

Mixed-effects models were constructed only for the relatively more frequent affective states: engagement/flow, boredom, confusion, curiosity, happiness. Though relatively less frequent, frustration was also included in the analyses due to the interesting between-study variability in effect size distributions for this state (see previous section). Separate models were constructed for each moderator in order to individually assess its predictive power. This resulted in 24 models (6 affective states \times 4 moderators), as shown in Table 4.

Significant models were discovered for engagement/flow, boredom, confusion, and frustration but not for curiosity and happiness, presumably due to the small sample size for these two states. An

examination of the model coefficients indicated that students were less likely to self report being in the engaged/flow state, but they were more likely to report being bored and frustrated. This suggests that the source of the affect judge (self vs. observers) can impact what is being reported. It might be the case that observers have difficulty in detecting boredom and frustration due to subtle and conflicting cues associated with natural displays of these affective states. An analysis of facial expressions of boredom and engagement/flow has indicated that, compared to states like confusion and delight, these affective states have more subtle facial markers (McDaniel et al., 2007). This would make it difficult for observers to detect these states. Frustration provides an interesting case study for emotion perception because natural frustration is sometimes expressed with brief half smiles (Hoque & Picard, 2011), which are often confused with emotions like delight (McDaniel et al., 2007).

Studies conducted in authentic learning contexts were associated with more engagement/flow and less boredom and frustration. This is an expected finding, because motivation and engagement are likely to be higher when the learning task is aligned with students' schoolwork to the extent that students perceive intrinsic value in the learning activities. In contrast, boredom was higher in

Table 4
Model Summaries and Standardized Coefficients of Mixed-Effects Models

Affective state and moderator	<i>k</i>	Coefficients			Heterogeneity		
		<i>B</i> (<i>SE</i>)	95% CI [<i>LL</i> , <i>UL</i>]	<i>Z</i>	<i>QE</i> ^c	<i>QM</i> ^d	<i>Tau</i> ²
Engagement/flow							
Self-report	19	−3.07 (0.58)	[−4.21, −1.92]	−5.26	97.1	27.72	1.16
Authentic context	19	2.87 (0.68)	[1.54, 4.20]	4.24	196.0	17.94	1.71
ALT used	19	1.82 (0.87)	[0.12, 3.51]	2.10	197.7	4.40	3.05
Min	19	0.00 (0.03)	[−0.06, 0.05]	−0.11	213.4	0.01	3.92
Boredom							
Self-report	21	1.69 (0.41)	[0.88, 2.49]	4.11	179.5	16.88	0.69
Authentic context	21	−1.79 (0.38)	[−2.53, −1.04]	−4.69	103.3	21.99	0.56
ALT used	21	−1.15 (0.47)	[−2.07, −0.24]	−2.47	142.0	6.08	0.98
Min	21	0.01 (0.02)	[−0.02, 0.04]	0.49	226.7	0.24	1.42
Confusion							
Self-report	20	−0.02 (0.61)	[−1.21, 1.17]	−0.03	227.2	0.00	1.48
Authentic context	20	0.18 (0.62)	[−1.03, 1.39]	0.29	207.5	0.09	1.47
ALT used	20	0.50 (0.56)	[−0.60, 1.59]	0.89	195.6	0.78	1.39
Min	20	0.03 (0.01)^a	[0.00, 0.05]	1.84	161.3	3.38	1.19
Curiosity							
Self-report	—	—	—	—	—	—	—
Authentic context	9	0.79 (0.95)	[−1.07, 2.66]	0.83	58.5	0.69	0.77
ALT used	9	0.83 (0.61)	[−0.38, 2.03]	1.35	53.3	1.81	0.63
Min	9	0.00 (0.02)	[−0.05, 0.05]	−0.03	72.4	0.00	0.86
Happiness							
Self-report	7	−0.39 (0.88)	[−2.13, 1.34]	−0.45	35.9	0.20	0.29
Authentic context	—	—	—	—	—	—	—
ALT used	—	—	—	—	—	—	—
Min	7	−0.03 (0.03)	[−0.08, 0.03]	−0.98	26.6	0.96	0.26
Frustration							
Self-report	21	7.04 (1.58)	[3.94, 10.15]	4.45	196.0	19.77	11.01
Authentic context	21	−6.70 (1.71)	[−10.06, −3.34]	−3.91	362.2	15.31	13.05
ALT used	21	−3.91 (2.10)^b	[−8.03, 0.21]	−1.86	367.1	3.47	21.17
Min	21	−0.04 (0.06)	[−0.15, 0.07]	−0.73	377.9	0.54	24.32

Note. Blank cells (—) for coefficients indicate that moderator was not included in the model due to data availability. Significant coefficients (at $p < .05$) are bolded at $p < .05$. *SE* = standard error; *CI* = confidence interval; *LL* = lower limit; *UL* = upper limit; ALT = advanced learning technology; Min = length of learning session in minutes.

^a $p = .07$. ^b $p = .06$. ^c *QE* = test for residual heterogeneity and was always significant at $p < .01$. ^d *QM* = test of moderators (significance mirrors significance of coefficients).

laboratory studies, presumably because students perceive little value in learning activities that are unrelated to their educational goals. Indeed, perceived value and goal congruence are considered to be key predictors of engagement and curiosity (Hidi, 2006; Pekrun et al., 2010).

The use of ALTs was also associated with lower rates of boredom and frustration coupled with an increase in engagement/flow. This might be attributed to the interactive nature of the technologies; their customized instruction via sophisticated student modeling; the presence of immediate, direct, and discriminating feedback; or to simple novelty effects. Finally, training time was a significant positive predictor of confusion, but it did not significantly predict any of the other states.

Percent reduction in heterogeneity for each moderator was computed as $100 * [(Tau^2_{no\ mod} - Tau^2_{with\ mod}) / Tau^2_{no\ mod}]$. Tau^2 with and without moderators were taken from Tables 2 and 3, respectively. The source of the affect judge (self vs. observers), the authenticity of the learning context, and the use of ALTs yielded heterogeneity reductions of 57%, 52%, and 18% averaged across flow/engagement, boredom, and frustration. Training time reduced heterogeneity of confusion by 14%.

Random-Effects Models on Pairwise Effect Sizes

The pairwise effect size distributions were also analyzed with random-effects models to complement the previous analyses on the overall effect sizes. Analyses were performed only for effect size distributions with at least 4 data points, in light of the power analysis discussed above. Standardized weighted mean pairwise effect sizes are shown in Table 5.

The results largely replicated the patterns with overall effect sizes. Engagement/flow was relatively more frequent than all the other states. Boredom was relatively less frequent than engagement/flow, not significantly different from confusion, and relatively more frequent than the other affective states. Confusion was relatively less frequent than engagement/flow; statistically equivalent to boredom, curiosity, and happiness; and relatively more frequent than the remaining states. The results were particularly

interesting for frustration. Though relatively less frequent than flow/engagement, boredom, and confusion, frustration was not significantly different than curiosity, happiness, and delight and was relatively more frequent than anger, anxiety, contempt, disgust, sadness, and surprise.

The average pairwise d_+ for all comparisons involving engagement/flow, boredom, confusion, curiosity, happiness, and frustration was 0.53 sigma, whereas anxiety, anger, disgust, contempt, sadness, delight, fear, and surprise yielded an average pairwise d_+ of -0.43 sigma. In summary, both the present results of the pairwise effect sizes and the previous results with overall effect sizes suggest that engagement/flow, boredom, confusion, curiosity, happiness, and frustration are the affective states that are more likely to occur during learning with technology, at least to the extent of the 24 studies included in this analysis.

General Discussion

The present paper focuses on identifying the moment-to-moment affective states that students' experience during learning with technology. Considerable variability in the distribution of affective states was expected due to differences in student populations, learning technologies, learning contexts, and methodologies used for affect monitoring. The pertinent question was whether a set of generalizable *learning-centered* affective states that transcend inherent differences among students, technologies, tasks, and methodologies could be identified. This question was addressed via a selective meta-analysis of 24 studies that systematically monitored the affective states of diverse samples of 1,740 students over the course of interaction with a variety of learning technologies. In this section, I discuss the major findings, align these findings with existing and emerging theoretical perspectives on affect and learning, consider implications of the findings for the design of ALTs, and discuss limitations and potential areas of research that are particularly promising for future work.

Table 5
Weighted Mean Effect Sizes (d_+) for Pairwise Effects

Affect 1	Affect 2													
	Ang	Anx	Bor	Con	Cmt	Cur	Del	Dis	Fear	Eng	Fru	Hap	Sad	Sur
Anger		-0.33	-0.94	-0.82	-0.20	-0.34	—	0.05	0.36	—	-0.80	-0.49	0.16	0.21
Anxiety			-0.91	-0.54	—	-0.68	-0.11	0.38	0.60	-1.24	-0.50	-0.07	0.49	0.22
Boredom				0.12	0.99	0.69	0.88	1.03	1.10	-1.49	0.48	0.43	0.98	1.08
Confusion					0.72	0.18	1.01	0.66	0.90	-1.76	0.29	0.07	0.72	1.38
Contempt						-0.14	—	0.21	—	—	-0.83	—	—	—
Curiosity							0.59	0.35	—	-0.87	0.00	—	—	0.93
Delight								—	—	-2.17	-0.23	—	—	0.55
Disgust									0.45	—	-0.82	-0.57	0.12	0.32
Fear										—	-1.08	-0.80	-0.24	-0.15
Engagement/flow											2.21	0.96	—	2.33
Frustration												0.18	0.86	0.78
Happiness													0.62	0.72
Sadness														0.13
Surprise														

Note. Significant d_+ s are bolded. Positive values indicate that Affect 1 is significantly greater than Affect 2 (reverse for negative values). For example, the Anger–Anxiety d_+ of -0.33 indicates that anger was less frequent than anxiety. Blank cells (—) indicate that there were insufficient data for analysis.

Overview of Major Findings

A number of conclusions can be drawn from the results of this analysis. The primary finding is that engagement/flow, boredom, confusion, curiosity, happiness, and frustration appear to be the set of discrete affective states that are relatively more frequent during learning with technology. The fact that engagement/flow was relatively highly frequent and its relative frequency increased with ALTs compared to less sophisticated learning environments suggests that the enhanced interactivity, personalized instruction, rapid feedback, and other features of ALTs have the intended effect of engaging students. Unfortunately, this result is somewhat tempered, as boredom was also relatively quiet frequent in several of the studies and more so with less interactive technologies.

It was also informative to discover that confusion was relatively frequent in several studies. Confusion, which is sometimes referred to as a *knowledge emotion* (Silvia, 2009) or an *epistemic emotion* (Pekrun, 2010), occurs when students experience impasses when processing new information that clashes with prior knowledge and exposes problematic misconceptions and erroneous mental models (D'Mello & Graesser, 2012; Graesser, Lu, et al., 2005; Piaget, 1952). Learning is expected to be positively impacted to the extent that students reason and problem solve to resolve impasses (VanLehn et al., 2003) and undergo a form of conceptual change by revising their mental models (Dole & Sinatra, 1998).

In addition to these three states (engagement/flow, confusion, and boredom), curiosity, happiness, and frustration were observed, albeit with lower relative frequency. Curiosity is expected to be prominent when interest and motivation in the learning activity are high (Hidi, 2006), when there is novelty (Berlyne, 1978; Silvia, 2009), and when students have some degree of choice over the specifics of the learning task (Cordova & Lepper, 1996). The learning contexts analyzed in this paper differed in the extent to which they supported these antecedents of curiosity, which is one hypothesis to explain the relatively lower occurrence of this state.

Frustration is a state that occurs when there is negative feedback, when important goals are blocked, when there is persistent failure, and when students are stuck and do not have an immediate plan to proceed (Burlison & Picard, 2004; D'Mello & Graesser, 2012; Stein & Levine, 1991). Frustration-inducing events occur quite frequently over the course of learning conceptually difficult topics or when students attempt to solve challenging problems. It might be the case that the relative incidence of frustration was not exceedingly high in the studies that were analyzed, because most ALTs do not let students persevere when they are stuck but offer hints, worked examples, and other opportunities to move the session forward. Indeed, the mixed-effects models revealed that frustration was lower in studies that used ALTs compared to simpler interfaces.

In contrast to frustration, happiness is expected to occur when there is positive feedback on a student action, when students get insights to resolve troublesome impasses, and when intermediate goals are attained (Stein & Levine, 1991). It is difficult to imagine students sustaining states of happiness during learning with technologies that actively advance the session by introducing new content and providing new problems to solve. This is because the new material can trigger an entirely different profile of affective states, thereby giving students little time to revel in their happi-

ness. Hence, the overall lower relative incidence of happiness is consistent with expectations.

The discussion so far has focused on the six affective states that were found to occur with notable frequency. However, negative evidence can also be quite compelling, and it is particularly informative that eight of the affective states analyzed were found to be relatively infrequent. These include contempt, anger, disgust, delight, anxiety, sadness, surprise, and fear. At first blush, the relatively low frequency of anxiety might be somewhat surprising, but it is important to note that the consequences of failure in the learning tasks were not severe. Anxiety is likely to be heightened during high-stakes learning tasks, such as preparing for an important exam or taking a standardized test (Zeidner, 2007). That being said, it is important to note that the levels of urgency of the tasks examined in this analysis are comparable to many real-world learning tasks (e.g., solving homework problems, writing a book report, reading the textbook), so the findings are applicable to tasks with low to moderate urgency.

Aside from anxiety and delight, the remaining six states that were relatively infrequent can be considered to be basic emotions (Ekman, 1992; Izard, 2007). This suggests that, with the exception of happiness (a basic emotion), the basic emotions might be less relevant in short, in-depth learning sessions with technology, at least when it comes to the 24 studies analyzed in this paper. This finding is intuitively plausible, because there is no adequate justification to expect the average student to experience persistent episodes of sadness and disgust when interacting with a reasonably well-engineered learning technology on a task of low to moderate urgency. The basic emotions might be more relevant during longer sessions (e.g., completing a dissertation), when stakes are high (e.g., studying for an important exam), or when there are interactions with peers and superiors, but this is entirely an empirical question. As it currently stands, it is a set of nonbasic affective states (and happiness) that play an active role when students complete short but focused learning tasks with technology.

Finally, the analysis identified four between-study moderators that were predictive of student affect. Methodological and contextual factors, composed of whether the affect labels were provided by the students themselves or by external observers and whether the study was conducted in more authentic classroom setting versus a lab, were generally more predictive than the use of ALTs and training time. The main finding was that engagement/flow was more frequent when an ALT was used, when the study was conducted in an authentic learning context, and when the affect judgments were provided by observers rather than the students themselves. In contrast, boredom and frustration were predominantly observed in laboratory studies, when simple computer interfaces were used in lieu of ALTs, and when affective states were measured via self-reports. Taken together, these findings highlight how study-level factors influence student affect. An important message is that the affective state distributions are mainly influenced by the activity, location, and source of the measurements. In particular, affective states are impacted by what activity the student is engaged in (ALT vs. simpler computer interface), where the measurement occurs (schools vs. classrooms), and who is performing the measurement (self vs. others). Indeed, affective states are highly situation dependent and contextually coupled, instead of being dispositional and context free.

Theoretical Implications

The finding that engagement/flow, boredom, confusion, curiosity, happiness, and frustration were the predominant affective states can be aligned with theories that specify how these states might arise from (a) appraisals of control and value of the learning activity (Pekrun, 2010; Pekrun et al., 2010), (b) appraisals of goal congruence and plan availability (Stein & Levine, 1991), and (c) impasse detection, impasse-resolution processes, and states of cognitive disequilibrium (Graesser, Lu, et al., 2005; Piaget, 1952; VanLehn et al., 2003). With the exception of the control-value theory, which attempts to be somewhat comprehensive, the other theories focus on only one or two affective states. Hence, a brief sketch of how the present findings can be aligned with an integrative account of these theories is provided below.

The control-value theory emphasizes how appraisals of the perceived value in and control over the learning activity predict the affective states that students experience (Pekrun, 2010; Pekrun et al., 2002). Students perform these appraisals at multiple instances during a learning session, and continual appraisals with respect to challenges and progress can trigger major changes in student affect. Engagement/flow is expected to be heightened when students see value in the learning activity and when there is an appropriate balance between skill and challenges, so that they have some control over the outcome of the activity (Pekrun et al., 2010). Like engagement, curiosity is expected to be increased when intrinsic motivation in the learning task is high, which in turn influences appraisals of value (Berlyne, 1978; Cordova & Lepper, 1996). In contrast, boredom occurs when value is low, when skills exceed challenges (too high control; Csikszentmihalyi, 1990), and when challenges exceed skills (too low control; Acee et al., 2010; Pekrun et al., 2010).

On the basis of appraisals of control and value, students are typically in a state of either (a) *engagement/flow*, when they pursue the superordinate goal of mastering the material in the learning environment, or (b) *disengagement (boredom)*, when they abandon pursuit of the superordinate learning goal. According to goal appraisal theories (Mandler, 1999; Stein & Levine, 1991), events that arise during the learning activity are constantly being appraised with respect to their congruence with the superordinate goal. Appraisals of these events, particularly with respect to goals, give rise to affective states. The arousal level (intense/weak) of the affective state is dependent upon how relevant the event is to the goal, whereas the valence (positive/negative) depends on whether the event is congruent or incongruent with respect to the goal (Mandler, 1984; Stein & Levine, 1991). Events that are consistent with the achievement of goals result in positive affective states, such as happiness, whereas outcomes that jeopardize goal achievement can result in negative affective states, such as frustration.

Confusion and frustration are states of considerable interest because they trigger subgoals associated with resolving goal-blocking events. Theories that emphasize the importance of interruptions, impasses, and cognitive disequilibrium to learning (Graesser, Lu, et al., 2005; Mandler, 1984, 1999; Siegler & Jenkins, 1989; VanLehn et al., 2003) posit that the students get confused when they are confronted with troublesome impasses and when they are uncertain about what to do next. The student can initiate a subgoal of effortful reasoning and problem solving to resolve the impasse and restore equilibrium. Equilibrium is re-

stored when the source of the discrepant information is discovered, the impasse is resolved, and the student reverts back into the state of engagement/flow (D'Mello & Graesser, 2012). Frustration occurs when the impasse cannot be resolved, when the student gets stuck, or when there is no available plan to resolve the goal-blocking event (Stein & Levine, 1991).

Implications for ALTs

The results have important implications for ALTs that do not currently model affective states as well as for next-generation systems that aim to be sensitive to students' affective states in addition to their cognitive states. On one hand, the fact that engagement/flow was the most (relatively) frequent affective state is promising for current learning technologies, especially for those that aspire to keep students focused; for example, by engaging in adaptive dialogue moves similar to human tutors (Graesser, Chipman, et al., 2005) or by incorporating narratives, seductive details, and increasing interactivity, as is the case with simulation and game-based environments (Sabourin et al., 2011). On the other hand, boredom was relatively frequent and curiosity was relatively infrequent (compared to boredom); this implies that there is still considerable room for improvement, especially if the goal is to increase motivation, interest, and engagement. This is an important goal, because active engagement is a prerequisite to the deployment of key cognitive and metacognitive processes, such as effortful problem solving, self-explanation, prior knowledge activation, planning, and inference generation. Hence, more basic research is needed to identify the factors that influence engagement and task persistence. Insights gleaned from such a research program can be integrated into ALTs so that they can increase curiosity and sustain long-term engagement over multiple sessions.

In a related but somewhat different vein, the present results suggest that the student models of next-generation ALTs should incorporate affective states in addition to cognitive states and knowledge levels. It might be possible to design educational technologies that increase positive affective states such as engagement, curiosity, and happiness, but it is highly unlikely that such systems will prevent the occurrence of boredom, frustration, and other negative affective states. These negative states can have crippling effects on task persistence, motivation, and learning gains (Craig et al., 2004; D'Mello & Graesser, 2011; Linnenbrink & Pintrich, 2002; Pekrun et al., 2010), so next-generation ALTs should incorporate mechanisms to intelligently handle the inevitable occurrence of negative affective states in a manner that is contextually constrained and dynamically adaptive to individual students. Although the research community is beginning to make considerable progress along this front (Arroyo et al., 2009; Conati & Maclaren, 2009; D'Mello et al., 2010; du Boulay et al., 2010; Forbes-Riley & Litman, 2009; Sabourin et al., 2011; Sazzad, AlZoubi, Calvo, & D'Mello, 2011), there is significant uncertainty associated with identifying which affective states to target, developing systems to detect these states, and implementing strategies to respond to the sensed states. The present analysis suggests that it might be advisable to initially focus on boredom, confusion, and frustration, because these were the relatively more frequent negative affective states across several studies and learning technologies.

The first step in developing an affect-sensitive ALT to respond to boredom, confusion, and frustration requires automated meth-

ods to detect affect. Previous work has indicated that it is possible to automatically classify these states by analyzing contextual information (e.g., whether negative vs. positive feedback was delivered, difficulty of the current problem) and interaction features (e.g., time taken for students to respond to a question, response verbosity; Baker et al., 2012; D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008), along with behavioral and physiological markers that have been extensively studied by researchers in the fields of nonverbal behavior and affective computing (see these reviews: Calvo & D'Mello, 2010; D'Mello & Kory, 2012; Pantic & Rothkrantz, 2003; Valstar, Mehu, Jiang, & Pantic, 2012; Zeng et al., 2009). In addition to detecting affect, understanding why a particular affective state was triggered is critical in order to facilitate more nuanced interventions to regulate affect. For example, there appears to be a curvilinear relationship between perceived control and boredom, in that boredom is frequent when control is too high (i.e., skill outweighs challenges; Csikszentmihalyi, 1990) but also when control is too low (challenges outweigh skill; Pekrun et al., 2010). A possible strategy to respond to boredom stemming from appraisals of low control is to decrease problem difficulty, but it would be beneficial to ramp up difficulty when boredom emerges from appraisals of high control. As this example illustrates, interventions that are not informed by the potential causes that underlie an affective state are less likely to be effective and might even be harmful if the intervention is misaligned with the cause of an affective state. In other words, the affective state might be a symptom of an underlying cause, be it a lack of motivation or a lack of knowledge, so it is important to understand the cause in order to effectively treat the symptom. Understanding the causal structure that gives rise to particular affective states and distinct manifestations of a single state (e.g., different forms of boredom) requires fine-grained monitoring of the sequence of system- and learner-generated events immediately prior to an affective episode. This information not only is important to engineer effective affect-sensitive interventions but also contributes to the basic research on understanding affect during learning. It is impossible to obtain this level of fine-grained information at a large scale in more traditional learning settings (e.g., a classroom) or in technologically light settings. However, ALTs provide an excellent forum to investigate affective experience in fine-grained detail, because they are specifically designed to be dynamically sensitive to individual students at fine-grained levels (i.e., microadaptivity or the inner loop à la VanLehn, 2006), and they usually keep meticulous logs of system- and learner-generated events that can be automatically mined.

Limitations and Resolutions

There are four noteworthy limitations with this analysis. First, the relatively small number of studies analyzed is of some concern. The fact that only 24 studies were considered can be attributed to the small number of available studies with usable data and the informal nature of the search. Inconsistent measurement models, extremely low sample sizes, and the failure to publish error estimates on affect proportions were additional factors that contributed to the elimination of a number of studies. There was sufficient diversity with respect to the learning technologies, subject domains, student characteristics, and affect judgment methodologies, thereby alleviating some of the generalizability concerns.

Nevertheless, it would be advisable to repeat this analysis as additional studies on affect and learning emerge in the literature.

The second limitation was that there was an imbalance in the number of affective states considered across studies. Some affective states were included in a large number of studies, and others were considered in only a handful of studies. For example, sadness, contempt, and fear were included in only three studies, and they were found to be relatively infrequent. On the other hand, surprise, which had the lowest d_+ , was included in 14 studies. These examples illustrate that it is not clear whether the relative incidence of an affective state was affected by the number of studies that included that state. Fortunately, d_+ s for each state were not significantly correlated with the number of studies with available data ($r = .179, p = .541$). This suggests that imbalance in the number of affective states did not significantly impact the results, although more data are needed before one can be confident in the effect size distributions of states that were not sufficiently represented in this set of studies.

The third limitation emerges from the fact that most of the affective states considered in the studies tended to be activating states (i.e., states with moderate to high arousal). These included positive activating states such as happiness, delight, and curiosity as well as negative activating states such as anger, confusion, and frustration. Sadness and boredom were the only two negative deactivating states (states with lower arousal), and positive deactivating states, such as calmness and relaxed, were missing entirely. This imbalance with respect to deactivating states makes it impossible to make any claims about the relative incidence of these states. It is therefore advisable that future research studies should expand the set of affective states in order to strike more of a balance between activating and deactivating states.

Previous research that has performed between-domain and within-domain comparisons on motivational orientations and academic emotions has indicated important domain effects (Bong, 2001; Goetz, Frenzel, Pekrun, Hall, & Lüdtke, 2007; Goetz, Pekrun, Hall, & Haag, 2006). For example, Goetz et al. (2006) found that affective responses did not generalize from one domain to another and that the degree of domain-specificity varied across states. The present study did not include domain as a potential moderator, because the sample size of 24 studies was not sufficiently large to accommodate the considerable diversity in learning domains. This is the fourth limitation, and it should be addressed with a larger sample of studies.

Concluding Remarks

This paper investigated one of the most fundamental issues in the burgeoning research area on affect during learning with technology. The purpose was to identify a set of discrete affective states that generalize across student populations, subject domains, learning contexts, learning technologies, and methods used to monitor affect. Learning was narrowly construed as a short but involved exchange between a student and some form of educational technology, and the focus was on fine-grained assessments of natural expressions of discrete affect.

The key take-home messages are that (a) affect distributions are consistent with a three-level hierarchy consisting of engagement/flow that was relatively frequent in all the studies; boredom, confusion, curiosity, happiness, and frustration, whose relatively

frequency varied across studies; and eight relatively infrequent affective states (contempt, anger, disgust, sadness, anxiety, delight, fear, and surprise); (b) the relative frequency of the affective states is mainly related to the source of the affect judgment and the authenticity of the learning context; (c) it was not the basic emotions that have dominated the scientific landscape for over a century but, with the exception of happiness, a set of nonbasic affective states that was more relevant in the learning sessions that were analyzed; and (d) researchers might consider targeting boredom, confusion, and frustration as negative affective states to detect and address with appropriate strategies in affect-sensitive ALTs.

In conclusion, this selective meta-analysis was intended to serve as an initial step toward organizing and synthesizing some of the emerging research on affect and learning. The hope is that it will serve as a launching point toward more basic research on the relative incidence, antecedents, dynamics, and consequences of the five nonbasic affective states (plus happiness) that appear to be relatively more frequent during learning with technology. More advances in basic research should inspire and challenge the ALTs of the future to include mechanisms that increase motivation and sustain persistence by embracing rather than ignoring affect.

References

References marked with an asterisk indicate studies included in the meta-analysis.

- Acee, T. W., Kim, H., Kim, H. J., Kim, J.-I., Chu, H.-N. R., Kim, M., . . . Wicker, F. W. (2010). Academic boredom in under- and over-challenging situations. *Contemporary Educational Psychology, 35*, 17–27. doi:10.1016/j.cedpsych.2009.08.002
- *Afzal, S., & Robinson, P. (2011). Natural affect data: Collection and annotation. In R. A. Calvo & S. K. D'Mello (Eds.), *New perspectives on affect and learning technologies* (pp. 55–70). New York, NY: Springer. doi:10.1007/978-1-4419-9625-1_5
- Ainley, M. (2008). Interest: A significant thread binding cognition and affect in the regulation of learning. *International Journal of Psychology, 43*, 17–18.
- Ainsworth, S. (2008). How do animations influence learning? In D. Robinson & G. Schraw (Eds.), *Current perspectives on cognition, learning, and instruction: Recent innovations in educational technology that facilitate student learning* (pp. 37–67). Charlotte, NC: Information Age.
- Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education, 19*, 105–154.
- *Alhothali, A. (2011). *Modeling user affect using interaction events* (Master's thesis). University of Waterloo, Waterloo, Ontario, Canada.
- Allison, P. D. (1999). *Multiple regression*. Thousand Oaks, CA: Pine Forge Press.
- Arroyo, I., Woolf, B., Cooper, D., Burleson, W., Muldner, K., & Christopherson, R. (2009). Emotion sensors go to school. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 17–24). Amsterdam, the Netherlands: IOS Press.
- Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students "game the system". In E. Dykstra-Erickson & M. Tscheligi (Eds.), *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (pp. 383–390). New York, NY: Association for Computing Machinery.
- Baker, R. J. S. D., D'Mello, S. K., Rodrigo, M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*, 223–241. doi:10.1016/j.ijhcs.2009.12.003
- *Baker, R. S. J. D., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., . . . Rossi, L. (2012). Towards sensor-free affect detection in cognitive tutor algebra. In K. Yacef, O. Zaïane, H. HersHKovitz, M. Yudelson, & J. Stamper (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 126–133). Retrieved from http://educationaldatamining.org/EDM2012/uploads/procs/Full_Papers/edm2012_full_1.pdf
- *Baker, R. S. J. D., Moore, G. R., Wagner, A. Z., Kalka, J., Salvi, A., Karabinos, M., . . . Yaron, D. (2011). The dynamics between student affect and behavior occurring outside of educational software. In S. D'Mello, A. Graesser, B. Schuller, & J. Martin (Eds.), *Proceedings of the fourth biannual International Conference on Affective Computing and Intelligent Interaction* (pp. 14–24). Berlin, Germany: Springer-Verlag.
- Baker, R., Rodrigo, M., & Xolocotzin, U. (2007). The dynamics of affective transitions in simulation problem-solving environments. In A. Paiva, P. Rui, & W. Rosalind (Eds.), *Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction* (pp. 666–677). Berlin, Germany: Springer.
- Beal, C. R., Arroyo, I. M., Cohen, P. R., & Woolf, B. P. (2010). Evaluation of AnimalWatch: An intelligent tutoring system for arithmetic and fractions. *Journal of Interactive Online Learning, 9*, 64–77.
- Berlyne, D. (1978). Curiosity in learning. *Motivation and Emotion, 2*, 97–175. doi:10.1007/BF00993037
- Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task value, and achievement goals. *Journal of Educational Psychology, 93*, 23–34. doi:10.1037/0022-0663.93.1.23
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley. doi:10.1002/9780470743386
- Burleson, W., & Picard, R. (2004, August). *Affective agents: Sustaining motivation to learn through failure and a state of "stuck."* Paper presented at the ITS 2004 Workshop Proceedings on Social and Emotional Intelligence in Learning Environments, Maceió, Brazil.
- Calvo, R. A., & D'Mello, S. K. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*, 18–37. doi:10.1109/T-AFFC.2010.1
- Calvo, R. A., & D'Mello, S. K. (Eds.). (2011). *New perspectives on affect and learning technologies*. New York, NY: Springer. doi:10.1007/978-1-4419-9625-1
- Chaouachi, M., & Frasson, C. (2010). Exploring the relationship between learner EEG mental engagement and affect. In J. Kay & V. Aleven (Eds.), *Proceedings of 10th International Conference on Intelligent Tutoring Systems* (pp. 291–293). Berlin, Germany: Springer.
- Clifford, M. (1988). Failure tolerance and academic risk-taking in ten- to twelve-year-old students. *British Journal of Educational Psychology, 58*, 15–27. doi:10.1111/j.2044-8279.1988.tb00875.x
- Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*, 101–129. doi:10.2307/3001666
- Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. doi:10.1037/0033-2909.112.1.155
- Conati, C., & McLaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction, 19*, 267–303. doi:10.1007/s11257-009-9062-8
- Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. In M. Bauer, P. Gmytrasiewicz, & J. Vassileva (Eds.), *Proceedings of the 8th International Conference on User Modeling* (pp. 137–147). Berlin, German: Springer.

- Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88, 715–730. doi:10.1037/0022-0663.88.4.715
- *Craig, S. D., D'Mello, S., Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with AutoTutor: Applying the facial action coding system to cognitive-affective states during learning. *Cognition & Emotion*, 22, 777–788. doi:10.1080/02699930701516759
- *Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media*, 29, 241–250. doi:10.1080/1358165042000283101
- Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety*. San Francisco, CA: Jossey-Bass.
- Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York, NY: Harper and Row.
- Daschmann, E. C., Goetz, T., & Stupnisky, R. H. (2011). Testing the predictors of boredom at school: Development and validation of the precursors to boredom scales. *British Journal of Educational Psychology*, 81, 421–440. doi:10.1348/000709910X526038
- Deci, E., & Ryan, R. (2002). The paradox of achievement: The harder you push, the worse it gets. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 61–87). Orlando, FL: Academic Press. doi:10.1016/B978-012064455-1/50007-5
- D'Mello, S., Craig, S., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18, 45–80. doi:10.1007/s11257-007-9037-6
- D'Mello, S., & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition & Emotion*, 25, 1299–1308. doi:10.1080/02699931.2011.613668
- D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22, 145–157. doi:10.1016/j.learninstruc.2011.10.001
- D'Mello, S., & Graesser, A. (in press). Confusion. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *Handbook of emotions and education*. New York, NY: Taylor & Francis. doi:10.1016/j.learninstruc.2012.05.003
- D'Mello, S., & Kory, J. (2012). Consistent but modest: Comparing multimodal and unimodal affect detection accuracies from 30 studies. In L.-P. Morency, D. Bohus, H. Aghajan, A. Nijholt, J. Cassell, & J. Epps (Eds.), *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 31–38). New York, NY: ACM.
- *D'Mello, S., Lehman, S., Pekrun, R., & Graesser, A. (2012). Confusion can be beneficial for learning. *Learning and Instruction*. Advance online publication. doi:10.1016/j.learninstruc.2012.05.003
- D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., . . . Graesser, A. (2010). A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In J. Kay & V. Aleven (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 245–254). Berlin, Germany: Springer.
- *D'Mello, S., & Mills, C. (in press). Emotions while writing about emotional and non-emotional topics. *Motivation and Emotion*.
- D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, 70, 377–398. doi:10.1016/j.ijhcs.2012.01.004
- Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist*, 33, 109–128. doi:10.1207/s15326985ep3302&3_5
- du Boulay, B., Avramides, K., Luckin, R., Martínez-Mirón, E., Méndez, G., & Carr, A. (2010). Towards systems that care: A conceptual framework based on motivation, metacognition and affect. *International Journal of Artificial Intelligence in Education*, 20, 197–229.
- Dweck, C. (2006). *Mindset*. New York, NY: Random House.
- Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6, 169–200. doi:10.1080/02699939208411068
- Forbes-Riley, K., & Litman, D. (2009). Adapting to student uncertainty improves tutoring dialogues. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 33–40). Amsterdam, the Netherlands: IOS Press.
- Forbes-Riley, K., & Litman, D. J. (2011a). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53, 1115–1136. doi:10.1016/j.specom.2011.02.006
- Forbes-Riley, K., & Litman, D. (2011b). When does disengagement correlate with learning in spoken dialog computer tutoring? In S. Bull & G. Biswas (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 81–89). Berlin, Germany: Springer.
- Goetz, T., Frenzel, A. C., Pekrun, R., Hall, N. C., & Lüdtke, O. (2007). Between- and within-domain relations of students' academic emotions. *Journal of Educational Psychology*, 99, 715–733. doi:10.1037/0022-0663.99.4.715
- Goetz, T., Pekrun, R., Hall, N., & Haag, L. (2006). Academic emotions from a social-cognitive perspective: Antecedents and domain specificity of students' affect in the context of Latin instruction. *British Journal of Educational Psychology*, 76, 289–308. doi:10.1348/000709905X42860
- Graesser, A., Chipman, P., Haynes, B., & Olney, A. (2005a). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48, 612–618. doi:10.1109/TE.2005.856149
- *Graesser, A., Chipman, P., King, B., McDaniel, B., & D'Mello, S. (2007). Emotions and learning with AutoTutor. In R. Luckin, K. Koedinger, & J. Greer (Eds.), *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (pp. 569–571). Amsterdam, the Netherlands: IOS Press.
- Graesser, A. C., Lu, S., Olde, B. A., Cooper-Pye, E., & Whitten, S. (2005). Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory & Cognition*, 33, 1235–1247. doi:10.3758/BF03193225
- *Graesser, A., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B. (2006). Detection of emotions during learning with AutoTutor. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 285–290). Austin, TX: Cognitive Science Society.
- Hammerstrøm, K., Wade, A., & Jørgensen, A.-M. K. (2010). *Searching for studies: A guide to information retrieval for Campbell Systematic Reviews*. Retrieved from <http://campbellcollaboration.org/lib/project/179/>
- Harley, J., Bouchet, F., & Azevedo, R. (2012). Measuring learners' co-occurring emotional responses during their interaction with a pedagogical agent in MetaTutor. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 40–45). Berlin, Germany: Springer-Verlag.
- Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6, 203–217. doi:10.1037/1082-989X.6.3.203
- Hidi, S. (2006). Interest: A unique motivational variable. *Educational Research Review*, 1, 69–82. doi:10.1016/j.edurev.2006.09.001
- Hoque, M. E., & Picard, R. W. (2011, March). *Acted vs. natural frustration and delight: Many people smile in natural frustration*. Paper presented at the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, Washington, DC.
- Hulleman, C. S., Durik, A. M., Schweigert, S. A., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology*, 100, 398–416. doi:10.1037/0022-0663.100.2.398

- Immordino-Yang, M. H., & Damasio, A. (2007). We feel, therefore we learn: The relevance of affective and social neuroscience to education. *Mind, Brain, and Education, 1*, 3–10. doi:10.1111/j.1751-228X.2007.00004.x
- Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science, 2*, 260–280. doi:10.1111/j.1745-6916.2007.00044.x
- Johnson, W., & Valente, L. (2008, July). *Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures*. Paper presented at the 20th National Artificial Intelligence Conference, Menlo Park, CA.
- Kapoor, A., Burleson, B., & Picard, R. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies, 65*, 724–736. doi:10.1016/j.ijhcs.2007.02.003
- Kappas, A. (2010). Smile when you read this, whether you like it or not: Conceptual challenges to affect detection. *IEEE Transactions on Affective Computing, 1*, 38–41. doi:ieeecomputersociety.org/10.1109/T-AFFC.2010.6
- *Lehman, B., D'Mello, S. K., Chauncey, A., Gross, M., Dobbins, A., Wallace, P., . . . Graesser, A. C. (2011). Inducing and tracking confusion with contradictions during critical thinking and scientific reasoning. In S. Bull & G. Biswas (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 171–178). New York, NY: Springer.
- *Lehman, B., D'Mello, S., & Person, N. (2008, June). *All alone with your emotions: An analysis of student emotions during effortful problem solving activities*. Workshop on Emotional and Cognitive Issues held in conjunction with the Ninth International Conference on Intelligent Tutoring Systems, Montreal, Quebec, Canada.
- Lench, H. C., Bench, S. W., & Flores, S. A. (2013). Searching for evidence, not a war: Reply to Lindquist, Siegel, Quigley, and Barrett (2013). *Psychological Bulletin, 139*, 264–268. doi:10.1037/a0029296
- Lepper, M., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 135–158). Orlando, FL: Academic Press. doi:10.1016/B978-012064455-1/50010-5
- Lindquist, K. A., Siegel, E. H., Quigley, K. S., & Barrett, L. F. (2013). The hundred-year emotion war: Are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011). *Psychological Bulletin, 139*, 255–263. doi:10.1037/a0029038
- Linnenbrink, E., & Pintrich, P. (2002). The role of motivational beliefs in conceptual change. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 115–135). Dordrecht, the Netherlands: Kluwer Academic. doi:10.1007/0-306-47637-1_6
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Litman, D., & Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication, 48*, 559–590. doi:10.1016/j.specom.2005.09.008
- Mandler, G. (1984). *Mind and body: Psychology of emotion and stress*. New York, NY: Norton.
- Mandler, G. (1999). Emotion. In B. M. Bly & D. E. Rumelhart (Eds.), *Cognitive science: Handbook of perception and cognition* (2nd ed., pp. 367–384). San Diego, CA: Academic Press.
- Mayer, R. (Ed.). (2005). *The Cambridge handbook of multimedia learning*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511816819
- McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., & Graesser, A. (2007). Facial features for affective state detection in learning environments. In D. McNamara & G. Trafton (Eds.), *Proceedings of the 29th annual meeting of the Cognitive Science Society* (pp. 467–472). Austin, TX: Cognitive Science Society.
- Meyer, D., & Turner, J. (2006). Re-conceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review, 18*, 377–390. doi:10.1007/s10648-006-9032-1
- *Mills, C., & D'Mello, S. K. (2012). Emotions during writing on topics that align or misalign with personal beliefs. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 638–639). Berlin, Germany: Springer-Verlag.
- Mota, S., & Picard, R. (2003, June). *Automated posture analysis for detecting learner's interest level*. Paper presented at the Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction, Madison, WI.
- Öhman, A. (2008). Fear and anxiety: Overlap and dissociation. In M. Lewis, J. M. Haviland-Jones, & L. Feldman Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 709–729). New York, NY: Guilford Press.
- Pantic, M., & Rothkrantz, L. (2003). Toward an affect-sensitive multimodal human-computer interaction [Review]. *Proceedings of the IEEE, 91*, 1370–1390. doi:10.1109/JPROC.2003.817122
- Parnes, S. J. (1975). *Aha! Insights into creative behavior*. Buffalo, NY: DOK.
- Pekrun, R. (2010). Academic emotions. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook: Vol. 2. Individual differences and cultural and contextual factors*. Washington, DC: American Psychological Association.
- Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. H. (2010). Boredom in achievement settings: Exploring control-value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology, 102*, 531–549. doi:10.1037/a0019243
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37*, 91–105. doi:10.1207/S15326985EP3702_4
- Pekrun, R., & Linnenbrink-Garcia, L. (Eds.). (in press). *Handbook of emotions and education*. New York, NY: Taylor & Francis.
- Piaget, J. (1952). *The origins of intelligence*. New York, NY: International University Press. doi:10.1037/11494-000
- Picard, R. (1997). *Affective computing*. Cambridge, MA: MIT Press.
- *Pour, P. A., Hussein, S., AlZoubi, O., D'Mello, S., & Calvo, R. (2010). The impact of system feedback on learners' affective and physiological states. In J. Kay & V. Aleven (Eds.), *Proceedings of 10th International Conference on Intelligent Tutoring Systems* (pp. 264–273). Berlin, Germany: Springer-Verlag.
- *Rodrigo, M., & Baker, R. (2011a). Comparing the incidence and persistence of learners' affect during interactions with different educational software packages. In R. A. Calvo & S. K. D'Mello (Eds.), *New perspectives on affect and learning technologies* (pp. 183–202). New York, NY: Springer. doi:10.1007/978-1-4419-9625-1_14
- *Rodrigo, M., & Baker, R. (2011b). Comparing learners' affect while using an intelligent tutor and an educational game. *Research and Practice in Technology Enhanced Learning, 6*, 43–66.
- Rosenberg, E. (1998). Levels of analysis and the organization of affect. *Review of General Psychology, 2*, 247–270. doi:10.1037/1089-2680.2.3.247
- *Sabourin, J., Mott, B., & Lester, J. (2011). Modeling learner affect with theoretically grounded dynamic bayesian networks. In S. D'Mello, A. Graesser, B. Schuller, & J. Martin (Eds.), *Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction* (pp. 286–295). Berlin, Germany: Springer-Verlag.
- *Sazzad, M. S., AlZoubi, O., Calvo, R. A., & D'Mello, S. K. (2011). Affect detection from multichannel physiology during learning. In S. Bull & G. Biswas (Eds.), *Proceedings of the 15th International Conference on*

- Artificial Intelligence in Education* (pp. 131–138). New York, NY: Springer.
- Schutz, P., & Pekrun, R. (Eds.). (2007). *Emotion in education*. San Diego, CA: Academic Press.
- Siegler, R., & Jenkins, E. (Eds.). (1989). *Strategy discovery and strategy generalization*. Hillsdale, NJ: Erlbaum.
- Silvia, P. J. (2009). Looking past pleasure: Anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. *Psychology of Aesthetics, Creativity, and the Arts*, 3, 48–51. doi:10.1037/a0014632
- Stein, N., & Levine, L. (1991). Making sense out of emotion. In W. Kessen, A. Ortony, & F. Kraik (Eds.), *Memories, thoughts, and emotions: Essays in honor of George Mandler* (pp. 295–322). Hillsdale, NJ: Erlbaum.
- *Strain, A., & D'Mello, S. (2011). Emotion regulation during learning. In S. Bull & G. Biswas (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 566–568). New York, NY: Springer.
- Valstar, M. F., Mehu, M., Jiang, B., & Pantic, M. (2012). Meta-analysis of the first facial expression recognition challenge. *IEEE Transactions on Systems, Man, and Cybernetics: Part B. Cybernetics*, 42, 966–979. doi:10.1109/TSMCB.2012.2200675
- VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16, 227–265.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46, 197–221. doi:10.1080/00461520.2011.611369
- VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21, 209–249. doi:10.1207/S1532690XCI2103_01
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1–48.
- Vuorela, M., & Nummenmaa, L. (2004). Experienced emotions, emotion regulation and student activity in a web-based learning environment. *European Journal of Psychology of Education*, 19, 423–436. doi:10.1007/BF03173219
- Woolf, B., Arroyo, I., Muldner, K., Burleson, W., Cooper, D., Dolan, R., & Christopherson, R. (2010). The effect of motivational learning companions on low achieving students and students with disabilities. In J. Kay & V. Aleven (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 327–337). Berlin, Germany: Springer.
- Zeidner, M. (2007). Test anxiety in educational contexts: Concepts, findings, and future directions. In P. Schutz & R. Pekrun (Eds.), *Emotions in education* (pp. 165–184). San Diego, CA: Academic Press. doi:10.1016/B978-012372545-5/50011-3
- Zeng, Z., Pantic, M., Roisman, G., & Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 39–58. doi:10.1109/TPAMI.2008.52

Received December 13, 2011

Revision received February 21, 2013

Accepted March 18, 2013 ■

Next-Generation Environments for Assessing and Promoting Complex Science Learning

Edys S. Quellmalz and Jodi L. Davenport
WestEd

Michael J. Timms
Australian Council for Educational Research, Melbourne,
Australia

George E. DeBoer
American Association for the Advancement of Science,
Washington, DC

Kevin A. Jordan, Chun-Wei Huang, and
Barbara C. Buckley
WestEd

How can assessments measure complex science learning? Although traditional, multiple-choice items can effectively measure declarative knowledge such as scientific facts or definitions, they are considered less well suited for providing evidence of science inquiry practices such as making observations or designing and conducting investigations. Thus, students who perform very proficiently in “science” as measured by static, conventional tests may have strong factual knowledge but little ability to apply this knowledge to conduct meaningful investigations. As technology has advanced, interactive, simulation-based assessments have the promise of capturing information about these more complex science practice skills. In the current study, we test whether interactive assessments may be more effective than traditional, static assessments at discriminating student proficiency across 3 types of science practices: (a) identifying principles (e.g., recognizing principles), (b) using principles (e.g., applying knowledge to make predictions and generate explanations), and (c) conducting inquiry (e.g., designing experiments). We explore 3 modalities of assessment: *static*, most similar to traditional items in which the system presents still images and does not respond to student actions, *active*, in which the system presents dynamic portrayals, such as animations, which students can observe and review, and *interactive*, in which the system depicts dynamic phenomena and responds to student actions. We use 3 analyses—a generalizability study, confirmatory factor analysis, and multidimensional item response theory—to evaluate how well each assessment modality can distinguish performance on these 3 types of science practices. The comparison of performance on static, active, and interactive items found that interactive assessments might be more effective than static assessments at discriminating student proficiencies for conducting inquiry.

Keywords: educational assessment, science education, multimedia, psychometrics, technology enhanced assessment

Multiple forces are converging to propel science testing into the digital age. Recent national science education frameworks and standards advocate a significant shift in focus to fewer, more integrated core ideas, deeper understanding of dynamic science

systems, and greater use of science inquiry practices. Federal accountability requirements call for science testing at the elementary, middle, and secondary grades. International, national, and state tests are turning to technology to improve the efficiency of large-scale testing and extend the standards assessed.

In view of these forces, science educators are concerned with the suitability of available assessments for measuring what students should know and be able to do in science. For example, the recent College Board *Standards for Science Success*, the National Research Council *Framework for K-12 Science Education*, and the draft *Next Generation Science Standards* recommend deeper learning such as understanding the fundamental nature and behavior of science systems, along with the inquiry practices scientists use to study system dynamics (College Board, 2009; National Research Council [NRC], 2012). Yet, most existing large-scale science accountability tests do not address the full range of valued science standards, particularly understanding science systems and inquiry practices (Darling-Hammond, 2010; Darling-Hammond & Pecheone, 2010). As a result, there is concern about the construct validity of science accountability tests, that is, that the prevalent

This article was published Online First September 9, 2013.

Edys S. Quellmalz and Jodi L. Davenport, WestEd, Redwood City, California; Michael J. Timms, Australian Council for Educational Research, Melbourne, Australia; George E. DeBoer, American Association for the Advancement of Science, Washington, DC; Kevin A. Jordan, Chun-Wei Huang, and Barbara C. Buckley, WestEd.

This material is based upon work supported by National Science Foundation Grants DRL-0814776 and DRL-0733345 awarded to WestEd (Edys S. Quellmalz, principal investigator).

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this article should be addressed to Edys S. Quellmalz, WestEd, 400 Seaport Court, Suite 222, Redwood City, CA 94063-2767. E-mail: equellm@wested.org

static, multiple-choice item format that relies on recognition of correct answers does not elicit evidence of integrated knowledge about science systems or abilities to conduct scientific inquiry (cf. Liu, Lee, & Linn, 2011; Quellmalz & Pellegrino, 2009; Smith, Wiser, Anderson, & Krajcik, 2006).

Technology-based science tests are being developed to address the perceived limitations of traditional tests. These interactive assessments are rapidly appearing in state, national, and international testing programs. For example, the 2009 National Assessment of Educational Progress (NAEP) included interactive computer tasks (ICTs). The 2006, 2009, and 2012 cycles of the Programme for International Student Assessment (PISA) include computer-based forms (Koomen, 2006; National Assessment Governing Board [NAGB], 2008). At the state level, Minnesota has an online science test with simulated laboratory experiments and investigations of phenomena such as weather or the solar system (Minnesota Department of Education, 2010). Utah is piloting science simulations in their assessments (King, 2011). Moreover, the state testing consortia are designing technology-enhanced items to test English Language Arts and Math common core standards, so it is likely that tests of the forthcoming *Next Generation Science Standards* will include innovative task and item formats. Further, results from the 2009 NAEP science ICTs showed that while most students performed well when making low-level observations from data, the majority of students performed poorly on complex assessment tasks involving multiple variables or strategic decision-making (National Center for Education Statistics [NCES], 2012). Though exam results reflect differences in student scores on interactive items, no studies have evaluated whether the interactive test modality is indeed more effective at distinguishing student proficiency on different science practice skills.

In this article, we report an empirical study to test whether interactive assessments may be more effective than traditional, static assessments at discriminating student proficiency across three types of science practices. We employed the three key science practice skills from the NAEP 2009 Framework; (a) identifying principles (e.g., stating or recognizing scientific principles), (b) using principles (e.g., applying knowledge to make predictions and generate explanations), and (c) conducting inquiry (e.g., designing, running and interpreting experiments). We developed assessments in three modalities with increasing levels of interactivity. The *static* modality is the most similar to traditional assessments in which the system presents still images and text and is not responsive to student actions. The *active* modality presents dynamic portrayals of phenomena, such as animations, which the student can observe and review. Finally, the *interactive* modality presents dynamic representations of the science phenomena and is responsive to student actions.

Why Use Simulations to Assess Science Practices?

The powerful capabilities of technology may hold the key to transforming the range of science knowledge and practices that can be assessed (Quellmalz & Haertel, 2004; Quellmalz & Pellegrino, 2009; Quellmalz et al., 2011). Scientists use physical, mathematical, and conceptual models as tools for asking questions, testing hypotheses, and communicating findings about natural and designed systems (Clement, 1989; Nersessian, 2008). Simulations and modeling tools dynamically represent the spatial, causal, and

temporal processes in science systems and permit active, virtual investigations of phenomena that are too big or small, fast or slow, or dangerous to be conducted in hands-on labs (de Jong, 2006; Lehrer, Schauble, Strom, & Pligge, 2001; Quellmalz & Pellegrino, 2009; Stewart, Carter, & Passmore, 2005). As simulations and models become “tools of the trade” in science, they also become important mechanisms for allowing students to demonstrate a range of science practices including making observations and designing and carrying out investigations.

Numerous studies that have illustrated the benefits of science simulations for student learning have also demonstrated the potential of simulations for assessment. Simulations can support the development of deeper understanding and better problem-solving skills in areas such as genetics, environmental science, and physics (cf. Adams et al., 2008; Buckley, Gobert, Horwitz, & O'Dwyer, 2010; Horwitz, Gobert, Buckley, & O'Dwyer, 2010; Krajcik, Marx, Blumenfeld, Soloway, & Fishman, 2000; Schwartz & Heiser, 2006; Zacharia, 2007). For instance, students using an aquatic ecosystem simulation or a collective simulation of multiple human body systems were able to demonstrate causal connections among the levels of these systems (Hmelo-Silver et al., 2008; Ioannidou et al., 2010; Slotta & Chi, 2006; Vattam et al., 2011). Using a computer-based, simulation tool that allowed students to create, test, and revise models helped students develop more robust and transferrable modeling skills than worksheet-based instruction (Papaevipridou, Constantinou, & Zacharia, 2007). Interactive assessment tasks that take advantage of the affordances of simulations have the potential to capture evidence of progress on the use of complex science practices and to transform how science is tested.

Designing Effective Assessments

Though simulations can provide compelling environments for evaluating a range of science practices, assessments are only as effective as their design. Research provides guidance regarding the identification of a scientifically appropriate context, the alignment between assessment tasks and learning objectives, and the minimization of extraneous cognitive processing. Below we summarize research related to these facets of effective assessment design and present a set of design principles that were used to guide the development of the assessments for the current study.

Selecting a Scientifically Appropriate Context

Taking Science to School and *Applying Cognitive Science to Education* recommend that rather than teaching and testing individual ideas and skills separately, knowledge and skills be taught and tested in the context of a larger investigation linked to a driving question. Currently, PISA, Trends in International Mathematics and Science Study (TIMSS), NAEP, and state tests administer problem-based sets of inquiry tasks set in authentic contexts. The reports recommend that assessments probe integrated knowledge structures (schema), contextualize items in meaningful tasks, and address not just declarative and procedural knowledge, but also measure schematic knowledge and strategic reasoning in problem solving and inquiry tasks. The *Framework for K-12 Science Education* and draft *Next Generation Science Standards* echo these research-based design principles.

The recommendations for assessments resonate with cognitive science research on expertise. Across academic and practical do-

mains, research on the development of expertise indicates that experts have acquired large, organized, interconnected knowledge structures, called schema, and well-honed, domain-specific problem-solving strategies (Bransford, Brown, & Cocking, 2000). Jacobsen characterized the schema of experts as “complex systems” mental models in contrast to the deterministic “clock-work” mental models of novices (Jacobson, 2001). Learning theory holds that the learning environments in which students acquire and demonstrate knowledge should represent contexts of use (Collins, Brown, & Newman, 1989; Simon, 1980). Critics of current testing practices cite the overemphasis on disconnected, decontextualized declarative and procedural knowledge in contrast to integrated schematic knowledge and strategic problem solving and inquiry (Linn & Eylon, 2011; Quellmalz & Pellegrino, 2009).

Aligning Assessment Tasks and Learning Objectives

If students have a deep understanding of a science system, they should both understand core concepts and be able to use their knowledge to make inferences and conduct scientific investigations. Thus, the challenge of science assessment is to develop tasks that do not simply tap into disconnected bits of declarative and procedural knowledge but that call for the schematic and strategic knowledge needed to reason about complex systems and engage in scientific inquiry practices.

The NRC report, *Knowing What Students Know*, integrated the learning research summarized in *How People Learn* with advances in measurement science to describe systematic test design frameworks. The evidence-centered assessment design framework provides a strategy for ensuring that assessments tasks are aligned with the learning objectives (Messick, 1994; Mislavy, Almond, & Lukas, 2003; Pellegrino, Chudowsky, & Glaser, 2001). The framework suggests a process that begins with a clear specification of learning objectives in what it refers to as a *student model*. For example, a part of the student model may be that “students will be able to design a controlled experiment to test a hypothesis.” Next, the framework specifies a *task model* that describes the features of the task and environment that will allow students to demonstrate they have mastered the skills specified in the student model. For instance, a simulation might allow students to set a number of variables and observe the results of trials run with the settings they selected. Finally, the *evidence model* specifies what student responses and scores would serve as evidence of proficiency.

For example, the rubric for the controlled experiment design task would assume student mastery if the student varied the single variable of interest and controlled the remaining variables across trials.

Cognitively principled assessment design for science begins with a *student model* derived from a theoretical framework of the kinds of knowledge structures and strategies students should demonstrate as evidence of their level of expertise. The model-based learning and national science education frameworks and standards identify the broad conceptual knowledge structures and inquiry practices deemed by the profession to be goals of science education (Achieve, 2012; American Association for the Advancement of Science [AAAS], 1993; College Board, 2009; NAGB, 2009; NRC, 2011). The science practices set forth for the 2009 Science NAEP guided the science inquiry targets specified in the student model for our study. Table 1 summarizes the science practices and their cognitive demands as specified in the 2009 NAEP Science Framework.

For the *task model*, we extracted tasks from learning research on inquiry. This work shows that students in kindergarten through eighth grade, with appropriate scaffolding, can engage in investigations, make hypotheses, gather evidence, design investigations, evaluate hypotheses in light of evidence, and build their conceptual understanding (Geier et al., 2008; Lehrer & Schauble, 2002; Metz, 2004). A number of studies provided evidence that such project-based experiences helped students learn scientific practices. Kolodner et al. (2003) found that middle school students who practiced inquiry in several project-based science units performed better on the inquiry tasks of scientific practice (as measured by performance assessments, Quellmalz, Schank, Hinojosa, & Padilla, 1999) than students from traditional classrooms. Moreover, all students, particularly English language learners, benefited greatly from inquiry-based science instruction that depended less on mastery of English than does decontextualized textbook knowledge or direct instruction by the teacher (O. Lee, 2002). Engaging students in active investigations allows students to demonstrate science inquiry skills and has been shown to increase conceptual understanding (cf. Dede, 2009; Geier et al., 2008; Kolodner et al., 2003; Lehrer & Schauble, 2002; Marx et al., 2004; Metz, 2004; Rivet & Krajcik, 2004).

Specifications of the *evidence models* were based on identifying the types of student responses within the simulation-based tasks that would serve as evidence of proficiency on science knowledge related

Table 1
2009 NAEP Science Practices

Science practice	Cognitive demand	Examples of skills related to practice
Identifying principles	Declarative knowledge, “Knowing that”	<ul style="list-style-type: none">• Knowing facts▪ Stating and recognizing science principles
Using principles	Schematic knowledge, “Knowing why”	<ul style="list-style-type: none">• Using patterns in observations▪ Making predictions▪ Creating explanations
Conducting inquiry	Procedural and strategic knowledge, “Knowing how and when”	<ul style="list-style-type: none">▪ Designing experiments▪ Testing predictions▪ Generating conclusions• Evaluating explanations

Note. NAEP = National Assessment of Educational Progress.

to three system model levels (components, interactions, and emergent system behavior) and the inquiry practices specified in the NAEP 2009 Framework. Rules were generated for scoring responses and summarizing them in order to report the proficiency levels.

Minimizing Extraneous Processing

Though visualizations and simulations have many affordances for learning, the additional information they present may also distract or overwhelm students. Multimedia learning researchers have examined the effects of pictorial and verbal stimuli in static, animated, and dynamic formats, as well as the effects of active versus passive learning enabled by degrees of learner control (Clark & Mayer, 2011; Lowe & Schnotz, 2007; Mayer, 2005b). Mayer's (2005a) *Cambridge Handbook of Multimedia Learning* and Clark and Mayer's recently updated book, *eLearning and the Science of Instruction* summarize multimedia research and offer principles for multimedia design (Clark & Mayer, 2011).

The majority of multimedia design principles address how to focus students' attention and minimize extraneous cognitive processing. Research suggests guiding attention by making the most important information salient and omitting irrelevant representations (cf. Betrancourt, 2005; Clark & Mayer, 2011). The use of visual cues such as text consistency, color, and arrows can help students map between representations and gain a deeper conceptual understanding (cf. Ainsworth, 2008; Kriz & Hegarty, 2007; Larkin & Simon, 1987; Lowe & Schnotz, 2008; Pedone, Hummel, & Holyoak, 2001).

Research specific to animations and interactive environments finds that the temporal nature of dynamic displays places increasing attention and memory demands on learners. To mitigate these additional demands, the research recommends (a) task-format alignment, (b) allowing for user control, (c) signaling upcoming changes, and (d) ensuring that the fidelity of the display is appropriate for the task. Task-format alignment suggests that dynamic or interactive features should be included only when they are required for the task. Animations are considered particularly useful for providing visualizations of dynamic phenomena that are not easily observable in real space and time scales (cf. plate tectonics, circulatory system, animal movement; Betrancourt, 2005; Kühl, Scheiter, Gerjets, & Edelmann, 2011). User control allows stu-

dents to pause, rewind, and replay dynamic presentations. Controlling the pace of the presentation can increase the likelihood that students will learn from and understand the display (cf. Lowe & Schnotz, 2008; Schwartz & Heiser, 2006). Signaling complex animations by giving cues such as "there will be three steps" and directly instructing students to reason through the components of systems increases student comprehension (Hegarty, 2004; Schwartz & Black, 1999; Tversky, Heiser, Lozano, MacKenzie, & Morrison, 2008). Mayer and Johnson (2008) found that redundancy of text in multimedia presentations may be beneficial when on-screen text is short, highlights the key action described in the narration, and is placed next to the portion of the graphic that it describes in order to highlight salient features of a multimedia presentation. Finally, the fidelity principle suggests that the complexity of a simulation should be appropriate for the learner outcomes. Rather than realistically portraying every detail of a system, it is more important to ensure that the most relevant parts of the system are easily discernible (cf. H. Lee, Plass, & Homer, 2006; van Merriënboer & Kester, 2005).

Animations become *interactive* simulations if learners can manipulate parameters as they generate hypotheses, test them, and see the outcomes, therefore taking advantage of technological capabilities well suited to conducting scientific inquiry. For example, Rieber, Tzeng, and Tribble (2004) found that students given graphical feedback during a simulation on laws of motion with short explanations far outperformed those given only textual information. Plass, Homer, and Hayward (2009) found that interactivity that allows for the manipulation of the content of a visualization, not just the timing and pacing, could improve learning outcomes compared to static materials. The authors suggest this is due to increased cognitive engagement (Plass et al., 2009).

From the previously cited bodies of literature we distilled design principles for designing next generation science assessments. The assessments for this study were designed according to recommendations for quality science assessments being made by science educators, cognitive scientists, and learning theorists. Table 2 summarizes the design principles that were used to ensure the use of scientifically appropriate contexts, the alignment between tasks and learning objectives, and the minimization of extraneous cognitive processing.

Table 2
Design Principles for Next Generation Science Assessments

Goal	Design principles
Identify scientifically appropriate context	<ul style="list-style-type: none"> ▪ Create rich environments that allow students to apply rich, interconnected knowledge ▪ Use authentic contexts to motivate assessment
Ensure alignment between tasks and learning objectives	<ul style="list-style-type: none"> ▪ Use evidence-centered design to ensure that tasks elicit evidence of proficiency for clearly specified learning goals ▪ Use task structures that tap into strategic and schematic knowledge
Minimize extraneous cognitive processing	<ul style="list-style-type: none"> ▪ Align the fidelity of the simulation with the task ▪ Eliminate interesting but task-irrelevant pictures and text ▪ Use visual cues to guide attention ▪ Ensure the task is appropriate for the multimedia in an item ▪ Allow users to control pace and replay of dynamic information ▪ Signal upcoming changes in animations

Method

To examine how well each modality of assessment distinguishes between the science practice constructs, we used three different analyses—a generalizability study, a Multitrait–Multimethod Confirmatory Factor Analysis, and a multidimensional Item Response Theory (IRT) model. Specifically, we investigated the following question: Do student responses on assessments in different modalities (static, active, and interactive) provide different information about students’ proficiencies on the three science practices: knowledge of science principles, use of science principles, and ability to conduct scientific inquiry?

Study Design

To answer our research question, we developed three parallel assessments in the context of the life science topic of ecosystems. Items in the three modalities were designed to test the same science practice constructs and to be comparable on all key stimulus and response features—except for dynamic representations of science phenomena and levels of interactivity, which varied across the static, active, and interactive modalities. We then analyzed our data using multiple psychometric and statistical techniques to determine whether the dynamic, active, and interactive assessments were better able to independently estimate student performance across the three science practices (*Identifying Principles, Using Principles, and Conducting Inquiry*).

Participants. A total of 1,836 students (910 female, 899 male, and 27 of unrecorded gender) from the classrooms of 22 middle school science teachers in 12 states participated in the study as part of normal classroom activities. Teachers received a stipend for the time needed to complete study activities (e.g., providing demographic information and enrolling students in the online learning management system). Due to absences, only 1,566 students (778 female, 776 male, 12 unknown) completed all three versions of the assessment. Thus, the total sample size included in the analyses was 1,566.

Materials. We applied the design principles described above to create three assessments in the context of ecosystems, each in the modality that reflected a differing level of interactivity: static, active, or interactive. The *static* assessment was designed to be similar to traditional, multiple-choice assessments (Figures 1 and 4). No part of the assessment was dynamic, that is, images and text were still, and the system was not responsive to student inputs. Visual representations for the static items were carefully chosen to be task relevant and minimize extraneous processing. The *active* assessment included items with dynamic displays such as animations. For instance, students could observe organisms interact and watch dynamic displays of experimental trials being run in a simulation environment. Figures 2 and 5 demonstrate active items that allow students to view dynamic ecosystems. Consistent with the principle to allow users to control the pace of dynamic information, the animations could be reviewed and paused as students observed predator-prey interactions. Finally, the *interactive* assessments shown in Figures 3 and 6 went beyond dynamic displays to permit student input that would result in a new screen reacting to the student input. Similar to the interactive items created for the 2009 NAEP Science Interactive Computer Tasks (prototypes of which were developed at WestEd), the interactive items created for the current study enabled learners to design experiments by manipulating parameters in a simulation, collect data by running simulations, and draw conclusions based on seeing the outcomes of their tests. The use of interactive feedback and simulations takes advantage of technological capabilities well suited to the science practice of conducting inquiry. Figure 6 shows an example of the interactive feedback. As students mouse over an organism, the name of that organism is highlighted in the legend, in the form of highlighting related information as students hover over organisms or a legend. This task feature reflects the design principle of using visual cues (color) to guide attention. Importantly, simulations allow students to design experiments and demonstrate their ability to change one variable at a time by setting sliders and to gather data by running the simulation and observing graphical and tabular

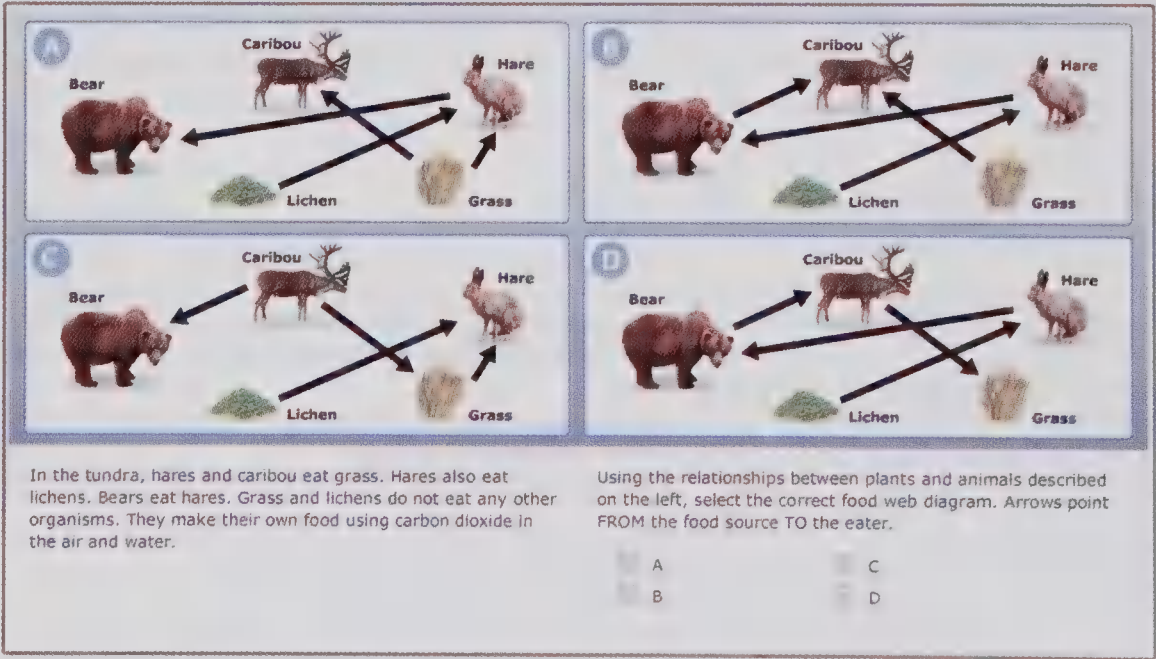


Figure 1. Ecosystem food web task: static modality.

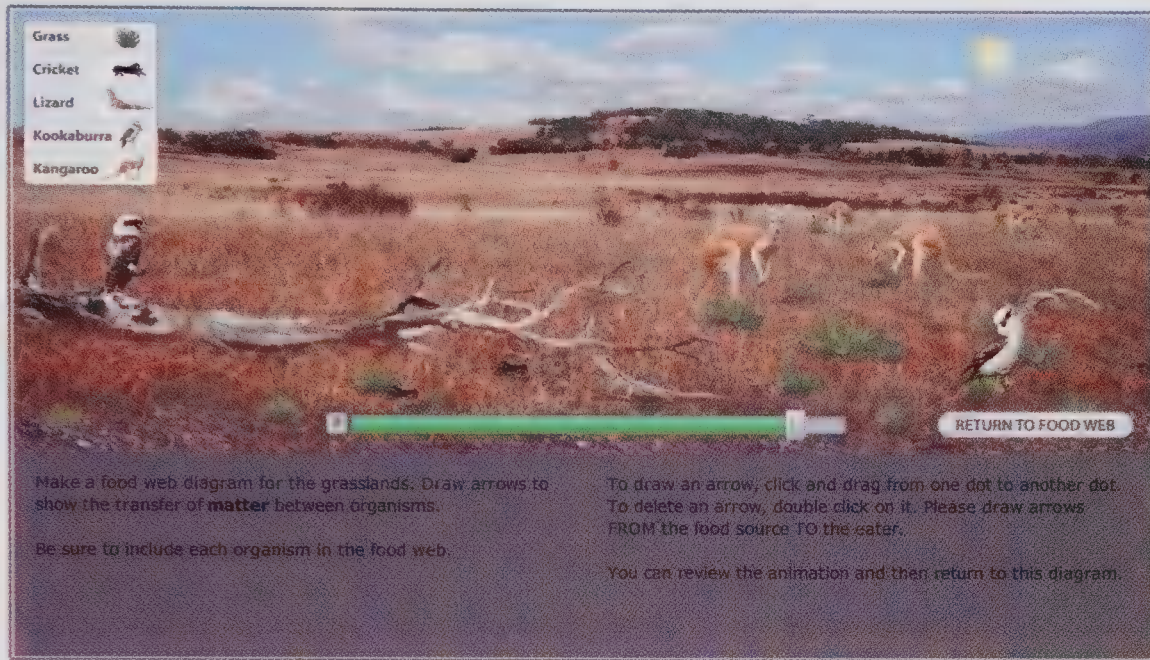


Figure 2. Ecosystem food web task: active modality.

output. Specifically, Figure 6 shows an item that allows students to set sliders to indicate the initial number of primary and secondary consumers and observe the effects of these settings on resulting population levels. To minimize extraneous processing, we used color as a visual cue to help students map between the values they selected and the outputs of the simulation.

The three modalities were designed to preserve the deep structure of each representation of the ecosystem model level (e.g., the same components, the same interactions among components, the same emergent population levels). The surface features, for example, the specific organisms, of each ecosystem varied. Each modality was presented in one of the three different ecosystem contexts (tundra, grasslands, and mountain lake). In the static modality, students viewed still images on the screen within a

tundra ecosystem. In the active modality, students viewed animations of a grasslands ecosystem but did not manipulate features or conduct active investigations. In the interactive modality, students identified and used ecosystem principles within a mountain lake ecosystem and conducted inquiry in tasks such as designing and running their own experiments on population levels. Each modality assessed the same three science practice constructs (Identifying Principles, Using Principles, and Conducting Inquiry).

For each of the three assessment modalities, evidence-centered design was used to create items that elicited evidence of proficiency on the three science inquiry constructs. Six items were designed to test the construct of *Identifying Principles*, six items were designed to test *Using Principles*, and 12 items were designed to test *Conducting Inquiry* for a total of 72 items in the

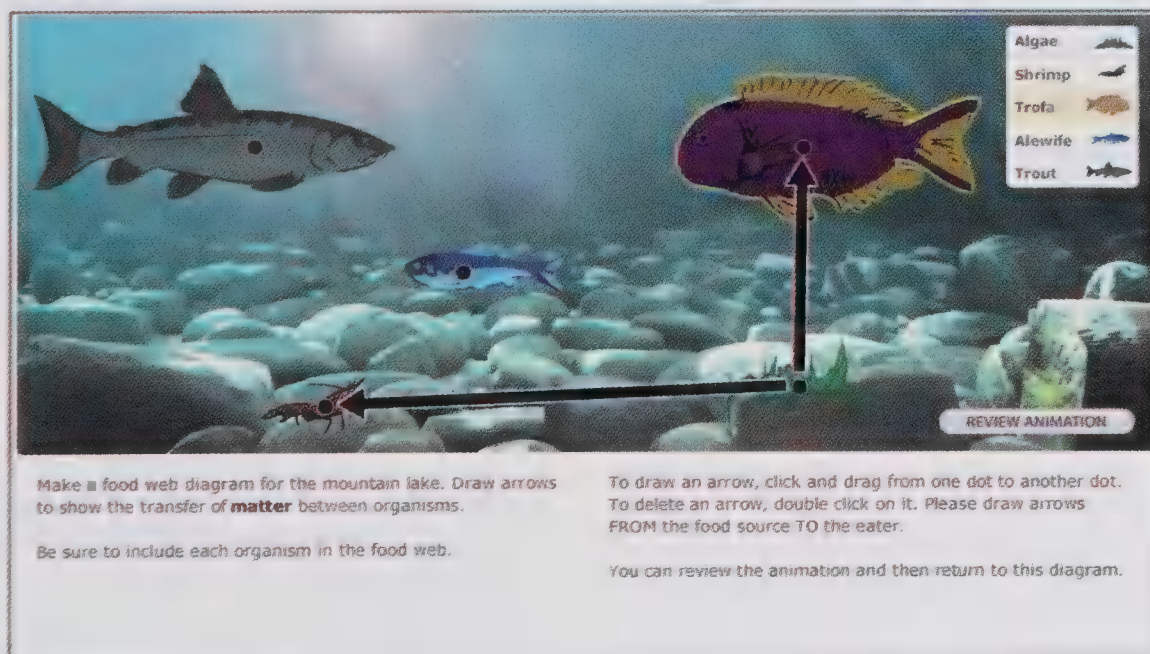


Figure 3. Ecosystem food web task: interactive modality.

context of ecosystems. A matrix was created with rows for each of the 72 items with specific statements related to the learning target in the context of ecosystems. For example, more specifically than assessing “Identifying principles,” the statement for a static item was “student is able to use descriptions and pictures of feeding relationships and correctly identify Yes/No whether each organism is a consumer.” The accompanying Active item statement could be, “student is able to observe predator prey interactions and identify Yes/No whether each organism is a consumer.”

To ensure our items reflected the design principle of scientifically appropriate content, we based our items on the interactive modality for ecosystems that had been developed and validated in prior research. The alignment of the interactive tasks with science practices for conducting inquiry had been established in prior projects. Similarly, the alignments of tasks for measuring, identifying, and using principles related to ecosystems had been established in prior research. Therefore, the representations of the components, interactions, and emergent population levels of ecosystems validated in our prior research formed the basis for modifying the ecosystem representations and assessment tasks for the three modalities (Quellmalz, Timms, & Buckley, 2010; Quellmalz et al., 2011). The existing ecosystem simulation environments and templates for ecosystems model levels and inquiry tasks were used to generate parallel sets of static, active, and interactive items.

Figures 1–3 illustrate the science constructs of *Identifying Principles* and *Using Principles* in a food web task in the different modalities. In the static modality, shown in Figure 1, students read text descriptions about the interactions of organisms (components) in an ecosystem, and students were asked to select a static image of the correct food web. The ability to correctly map the written description to a graphical food web is a component of the practice *Identifying Principles*. In the active modality, Figure 2, students observed an animation of organisms in an ecosystem and used their knowledge of the principles of organism roles (consumers, producers) to identify the roles of the organisms and then draw a food web diagram depicting the flow of energy and matter through

an ecosystem. Students could replay the animation. In the interactive modality, students observed the animation of an ecosystem and could take advantage of additional interactivity that used highlighting on demand to cue the connection between the names and pictures of the organisms. The skill of using observations to generate explanations (in the form of creating a food web diagram) was aligned with the practice *Using Principles*.

The potential confounding effects of nesting the assessments in a modality within a specific ecosystem (tundra, grasslands) were minimized by both maintaining the same structure of tasks and items for each modality expressed in each ecosystem and by focusing on assessment of science practices, not on knowledge of features of specific organisms or interactions that would differ between ecosystems.

Figures 4–6 illustrate items designed to test the science practice construct *Conducting Inquiry* in the three different modalities. In the static modality, students viewed the outcomes of an investigation, and students were asked to select an appropriate evaluation of the design. In the active modality, the student evaluated the design of an investigation after watching an animation of data being generated based on the design. In this active modality, students did not select inputs for the simulation, they only watched the simulation run. Finally, in the interactive modality, students designed their own investigations by setting the inputs for the simulation, running the simulation, observing the data tables and graphs being populated, and saving their own trials.

Design and procedures. All items were administered online, and data were collected using the SimScientists Learning Management System (Quellmalz, Timms, Silberglitt, & Buckley, 2012). The learning management system allows teachers to check on individual students and researchers to download de-identified student data. Initial construct validity of the items was examined by expert reviews and cognitive labs to confirm that the tasks and items were eliciting the intended practices.

Expert reviews. At two points in the process of developing the parallel item sets, three experts from the American Association for the

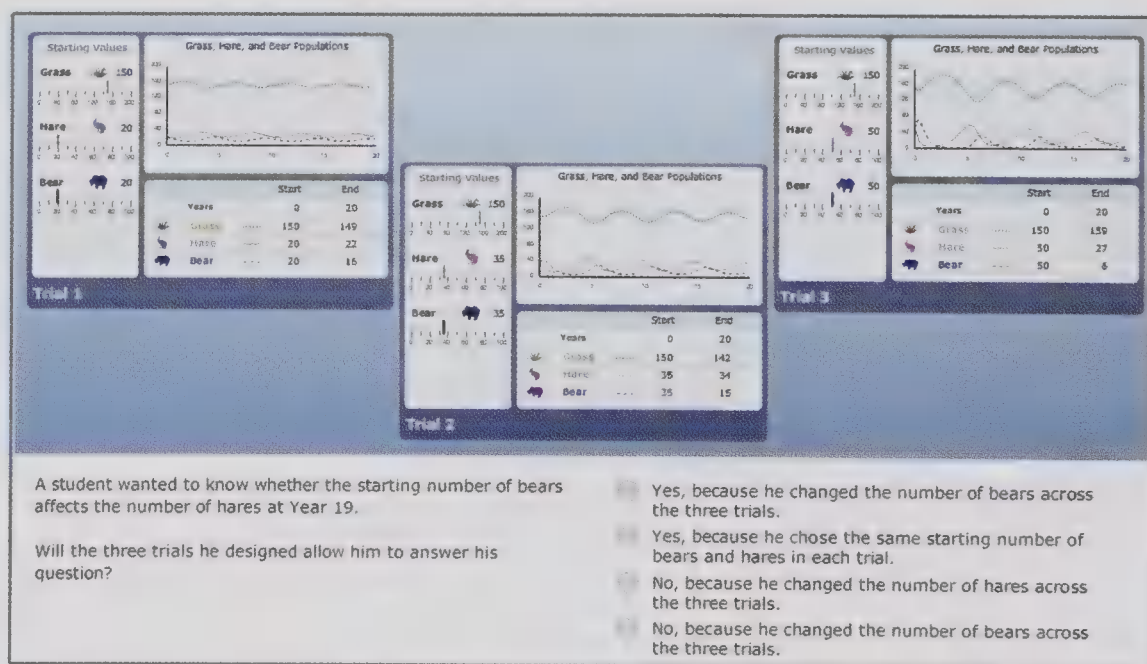


Figure 4. Investigation design task: static modality.

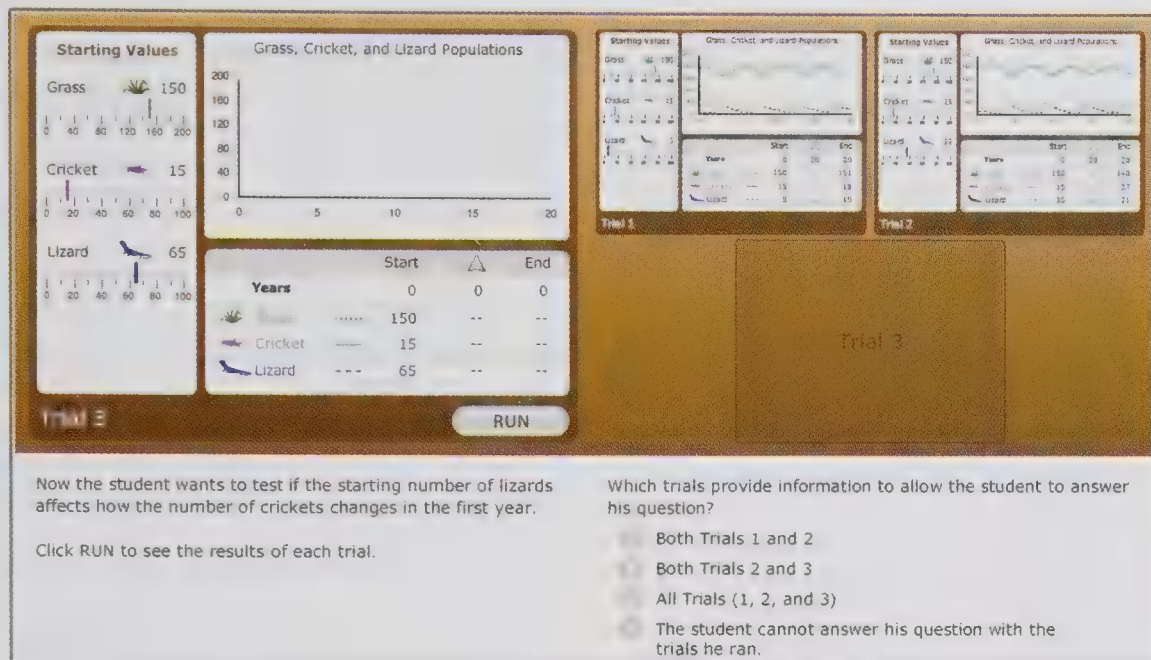


Figure 5. Investigation design task: active modality.

Advancement of Science (AAAS) independently reviewed the items and judged if each item was aligned with one of the targeted science practices of Identifying Principles, Using Principles, or Conducting Inquiry. These experts also verified that the items were scientifically accurate, grade-level appropriate, usable, and comparable across the static, active, and interactive versions. Initially, AAAS staff reviewed the storyboards of draft items and provided detailed comments and feedback. An additional iteration of review and revision was carried out with the programmed items to ensure the final items remained aligned with targeted science inquiry practices.

Cognitive labs. Ten students participated in think-aloud studies to determine if the items elicited the targeted science inquiry practices. Each student completed all three forms of the assessments, one in each modality (static, active, interactive). As students completed the

assessments, they “thought aloud” by saying everything they were thinking while screen capture software recorded students’ verbalizations and actions on the screen and researchers coded whether the items elicited the targeted construct. The think-aloud studies had two goals: (a) to ensure the usability of the assessments as deployed and (b) to provide evidence of construct validity by determining that the questions were eliciting student thinking and reasoning about the intended science practice constructs. To ensure the items would be usable in the field test, researchers took detailed notes of usability issues that arose (e.g., navigation, difficulty running experimental trials) and modified the items to address these issues. To examine the items’ construct validity, the observing researcher coded whether the item prompted student thinking related to the targeted science practice constructs. Table 3 summarizes the percentage of items in the assess-

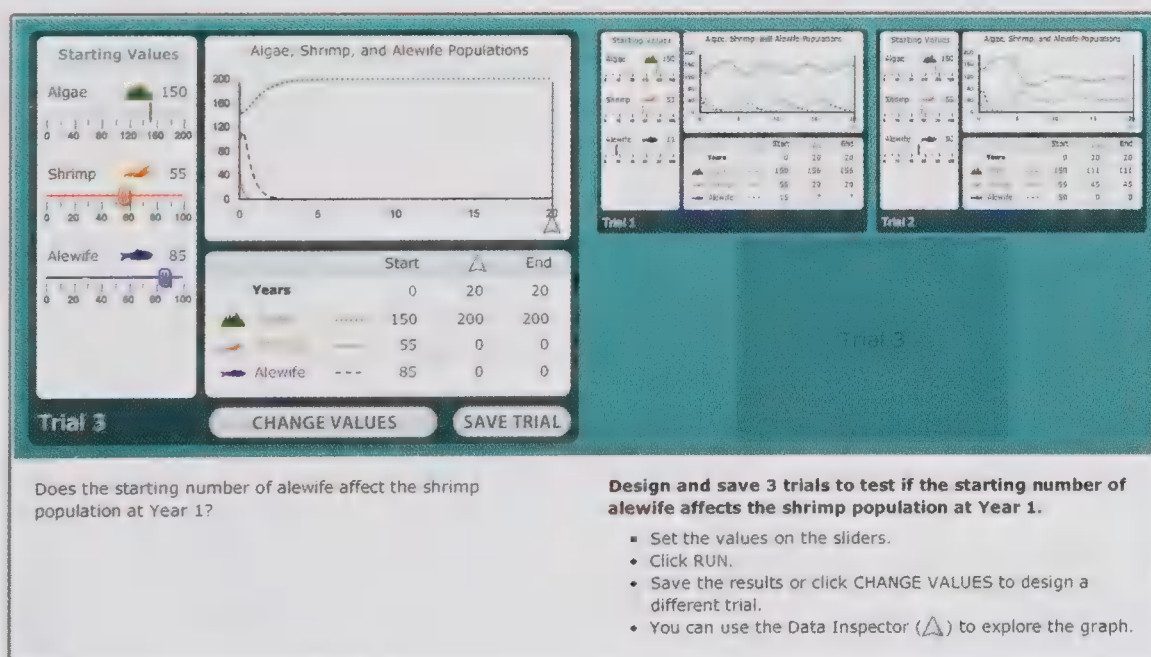


Figure 6. Investigation design task: interactive modality.

Table 3
Percentage of Items Judged to Elicit Their Intended Construct Targets

Modality	Identifying principles	Using principles	Conducting inquiry
Static %	100	97	98
Active %	98	98	100
Interactive %	98	98	98

ments judged to elicit their intended construct targets. These data provided one form of evidence that the items were aligned with their intended content and inquiry targets.

Procedures. The study used a within-subjects design, as all participating students took all three modalities of the ecosystems assessments, that is, one period of static items, one period of active items, and one period of interactive items. The assessments were given in three consecutive sessions, and the order of the sessions was fully counterbalanced at the class level across the six possible sequences of assessments: static (S), active (A), and interactive (I; SIA, ISA, IAS, AIS, ASI). Table 4 shows the number of items in each test session.

Participating teachers received a detailed "Teacher Guide" that outlined step-by-step the processes and procedures for study activities, and teachers were able to view online movies that demonstrated how to carryout the online processes and procedures necessary for participation. During the study, students either used laptops in their science classrooms or went to the school computer lab.

Analyses

In order to answer the research question, we used three different analyses to compare how well each assessment modality measured the three science practice constructs: (a) a generalizability study (G-study), (b) a Multitrait–Multimethod Confirmatory Factor Analysis (MTMM/CFA) and (c) a multidimensional Item Response Theory (MIRT) model. These three statistical methods employ models with successively stronger statistical assumptions. A G-study treats items as randomly sampled from a universe of potential items. It makes virtually no statistical assumptions beyond minimal assumptions that certain error components are uncorrelated. The MTMM/CFA imposes a specific theoretical model to test predictions about patterns of intercorrelations among the nine constructs (three science practice constructs by three assessment modalities). It is at a higher level of aggregation, moving from the item level to the level of nine composite scores that were formed according to the *a priori* framework underlying the instrument. The MIRT model returns to the item level but

imposes much stronger assumptions about the functional form of item characteristics as well as assumptions about the grouping of items into scales. More information for each modeling approach is described below.

Generalizability study. Generalizability analysis was chosen as the first analytic method. Generalizability studies (G-studies) extend the reliability analyses of classical test theory in powerful ways, making possible the quantitative investigation of multiple sources of error in a data set. Of particular relevance for the present study, multivariate G-studies enable simultaneous investigation of measures tapping multiple constructs, accounting for sources of correlated error across constructs. In the studies conducted here, G-studies indicate the magnitude of error variance components attributable to items as well as the interaction of persons by items plus residual variance.

The multivariate G-study analysis was conducted using the mGENOVA computer program (Brennan, 2001). Given that there were three ecosystem contexts for tasks (Grasslands, Tundra and Mountain Lake) corresponding to three item modalities (static, active, and interactive) and that each task consisted of items addressing three different science practice constructs (identifying, using, conducting), the analysis was a multivariate G-study treating the nine modalities by construct combinations as nine separate constructs. In the mGENOVA terminology, this is a ($p \bullet \times i^o$) design. The p facet represents persons (students). The solid circle means that the same students responded across all nine constructs. The i represents items. The open circle means that the items were nested within each of the nine constructs. This $p \bullet \times i^o$ notation represents the univariate design for each of the nine separate constructs. These constructs were specified and treated as a multivariate fixed facet in mGENOVA.

Multitrait–multimethod confirmatory factor analysis. The second type of analysis selected to answer the research question was a multitrait–multimethod CFA analysis (Campbell & Fiske, 1959; Loehlin, 1998), which attempts to separate out the true variance on measured traits (the underlying constructs being measured) from the variance that is due to the method of measurement (the modality). It is well-suited to this study because the same traits/constructs (the three science practices) were measured with three different methods (the static, active, and interactive modalities). The resulting correlations among the different measurements were then arranged in a multitrait–multimethod matrix in order to assess the *convergent validity*, the tendency for different measurement methods to converge on the same trait/construct, and the *discriminant validity*, the ability to discriminate among different traits/constructs.

Table 4
Number of Items in Each Test Session by Modality Format and Science Practice Construct

Science practice constructs	Test Session A static modality	Test Session B active modality	Test Session C interactive modality	Total items
Identifying science principles	6	6	6	18
Using science principles	6	6	6	18
Conducting science inquiry	12	12	12	36
Total items	24	24	24	72

This analysis was used to test the following two hypotheses that stem from the research question:

1. The factor loadings for method (assessment modality) are higher than for trait (science practice constructs).
2. The factor loadings from assessment modality to the three science practice constructs are less for the interactive modality than for the static and active modality.

The first hypothesis reflects our belief that the assessment modality is important in the assessment of science practices. If the factor loadings on the methods (assessment modality) are higher than those for the trait (science practice constructs), it tells us that the student scores are determined more by the assessment modality than by the science practice constructs being assessed (this is what we expect). If, on the other hand, the factor loadings are higher on the trait, it tells us that the science practice constructs are more important in determining student scores and the assessment modality is less important.

The second hypothesis examines which assessment modality measures more distinct science practice constructs. Higher correlations among science practice constructs for an assessment modality result in stronger factor loadings. This would indicate that the traits are not being measured distinctly under a certain assessment modality. We expect that the interactive modality measures the science practice constructs more distinctly than the other two modalities.

The multitrait-multimethod analysis used the same data set of 1,566 complete responses as was used for the G-study analyses. However, instead of using the responses to 72 individual items, nine composite scores were computed and used in the CFA. This was to simplify the analysis and the interpretation and to reduce the possibility of the estimation not being able to converge during analysis. The nine composite scores were simply the sum of the six items for *Identifying Principles*, the six items for *Using Principles* items, and the sum of 12 items for *Conducting Inquiry* items, computed for each of the three modalities (static, active, and interactive). The analysis was conducted following the procedures discussed on pp. 101–105 in Loehlin (1998). The Mplus¹ computer program was used for this analysis.

Multidimensional item response theory model. The third type of analysis used a multidimensional IRT model to evaluate how well each assessment modality (static, active, or interactive) was able to measure and separate student performances on the three sciences practice constructs. IRT models are probabilistic models in which item difficulty (a test item's underlying difficulty based on the proportion of a given sample that responded correctly) and person measure (a person's underlying competence, based on the proportion of items completed correctly) are simul-

Table 5
Estimated Correlations Among the Three Science Practices for the Static Mode (Tundra)

Science practice	Identifying	Using	Conducting
Identifying	1	0.92	0.80
Using		1	0.91
Conducting			1

Table 6
Estimated Correlations Among the Three Science Practice Constructs for the Active Mode (Grasslands)

Science practice	Identifying	Using	Conducting
Identifying	1	0.80	0.80
Using		1	1
Conducting			1

taneously estimated. The result is a scale on which both persons and items are mapped onto the theoretical latent traits, which in this case are the science practices constructs. The fact that IRT scores' accuracy and precision can be quantified makes this a suitable analytic method in this study to determine how well each of the three modalities of assessment measure the three science practice constructs.

The ACER ConQuest² generalized item response modeling program was used to run a multidimensional logistic model analysis that modeled the three science practice constructs separately for each of the three modalities of assessment. This allowed the correlations among the three science practice constructs to be estimated and the reliability of the measurement of each of the practice constructs to be quantified for the three modalities of assessment.

Results

G-study. We first summarize the estimated correlations among the three science practice constructs for each of the three modalities of assessment (see Tables 5–7). The estimated correlations in the G-study are corrected for attenuation due to unreliability, that is, estimated correlations among universe scores (true scores) for the nine constructs.³

These correlation coefficients can be used to examine which modality of assessment appears to measure more distinct science practice constructs. While we expect there to be a positive correlation among the three individual science practice constructs within the assessment modality (because they are related elements of the overall set of science practices, if they are clearly observable skills), the correlations should not be too high if the assessment is designed to measure multiple distinct constructs.

The results suggest that while similar patterns (correlations among science practice constructs) are found between assessment modes, the results in Table 7 show lower correlations between the constructs when tested by the interactive modality. The correlation between identifying and using for the interactive modality is 0.82 versus 0.92 and 0.80 for the static modality and the active modality, respectively. Using one minus the correlation as a measure of dissimilarity (it also represents the proportion of variance unique to each measure), we can see that moving from 0.82 (interactive) to 0.92 (static) represents a reduction in the variation unique to

¹ <http://www.statmodel.com/>

² ACER ConQuest: Generalized Item Response Modeling Software published and distributed by the Australian Council for Educational Research.

³ For this reason, the conventional methods to test the difference between observed correlations would not be appropriate.

Table 7
*Estimated Correlations Among the Three Science Practice
 Constructs for the Interactive Mode (Mountain Lake)*

Science practice	Identifying	Using	Conducting
Identifying	1	0.82	0.72
Using		1	0.84
Conducting			1

each modality from 18% to 8%, about 56% (10/18) reduction. This indicates that the correlation between identifying and using is more profound for the static modality than the interactive (or active) modality. Similarly, the correlation between identifying and conducting or the correlation between using and conducting is more apparent for the static/active modality than the interactive modality. Overall, the results suggest that the interactive modality measured the science practice constructs more distinctly than the other two modalities.

Table 8 summarizes the estimated variance components and G-coefficient by construct and indicates the percentage of variance attributable to each effect specified in the design. For example, under the static assessment modality for the *Identifying Principles* construct, about 20% of variance is contributed by persons (students), about 16% is from items, and 64% (the majority of variance) is accounted for by the person and item interaction. In general, this pattern holds for each construct. Note that these percentages shown in Table 8 are typical. A large percentage of variance for the person by item interaction is unsurprising as these percentages refer to single items, and a test with a single item is not expected to be reliable. In practice, items are always combined into a scale, and the person by item variance is divided by the number of items included in the scale (so more items would yield less variance for the person by item interaction).

The bottom row of Table 8 shows the G-coefficient estimate (a reliability-like coefficient) for each science practice construct measured under each of the three assessment modalities. The G-coefficients for the *Identifying* and *Using* constructs are based on a six-item test within each modality, whereas the coefficients for *Conducting* are based on a 12-item test (of multiple component skills) within each modality. There is little difference between the

G-coefficients for the static, active, and interactive modes in measuring the *Identifying* and *Using* constructs, but there is a noticeable difference for *Conducting*. The results suggest that the interactive mode produced a higher reliability (.79) than static (.68) or active (.66). This suggests the interactive modality produced a more reliable measure for *Conducting* than did the other two modalities.

Multitrait-multimethod confirmatory factor analysis.

Figure 7 shows the path model for the Multitrait-multimethod CFA. In the diagram, the top three circles represent the construct factors included in the model for the three science practices (identify, use and conduct). The bottom three circles represent the modality factors for the three assessment modalities (static, active, and interactive). The nine rectangles represent the sets of items. The first three on the left represent the items that targeted the *Identifying Principles* (identify) science practice, and from left to right they represent the static, active, and interactive item sets. The middle three rectangles represent the *Using Principles* (use) items, again with static, active, and interactive modalities running from left to right. The final three rectangles on the right represent the *Conducting Inquiry* (conduct) items, arranged yet again with static, active, and interactive modalities running from left to right. The straight lines represent the loadings of factors onto items, with the standardized value represented by the figure against the line. These factor loadings are summarized in Tables 9 and 10 below. The curved arrows represent the correlations between factors, with the value on each arc.

Our study question can be answered by looking at Tables 9 and 10. The factor loadings of assessment modalities to science practice items (see Table 10) are generally higher than the factor loadings of science practice factors (see Table 9). The findings indicate that the scores are determined more by the assessment modalities than by the science practice constructs. That is, consistent with our hypothesis, the modality of assessment does have an impact on how well items are able to draw out students' knowledge and skills in the three science practices.

By further looking at the factor loadings in Table 10, we also find that the factor loadings from the test modality to the three constructs are slightly less for the interactive modality (.698 on average) than for the static (.732 on average) and active modalities (.719 on average). This is supportive of our hypothesis that the task

Table 8
Estimated Variance Components and G-Coefficient by Construct

Effect	Identifying			Using			Conducting		
	Static	Active	Interactive	Static	Active	Interactive	Static	Active	Interactive
Persons									
Estimated variance	0.04	0.05	0.05	0.04	0.03	0.03	0.03	0.03	0.05
Percentage of total	19.62	22.27	20.67	16.48	14.83	14.79	11.97	11.59	18.11
Items									
Estimated variance	0.04	0.04	0.04	0.01	0.02	0.03	0.05	0.04	0.06
Percentage of total	15.93	14.64	15.15	5.96	7.17	12.30	19.92	16.33	22.49
Person \times Item interaction									
Estimated variance	0.15	0.15	0.16	0.17	0.17	0.17	0.17	0.17	0.15
Percentage of total	64.45	63.09	64.18	77.56	78.00	72.91	68.11	72.08	59.40
G-coefficient	0.65	0.68	0.66	0.56	0.53	0.55	0.68	0.66	0.79

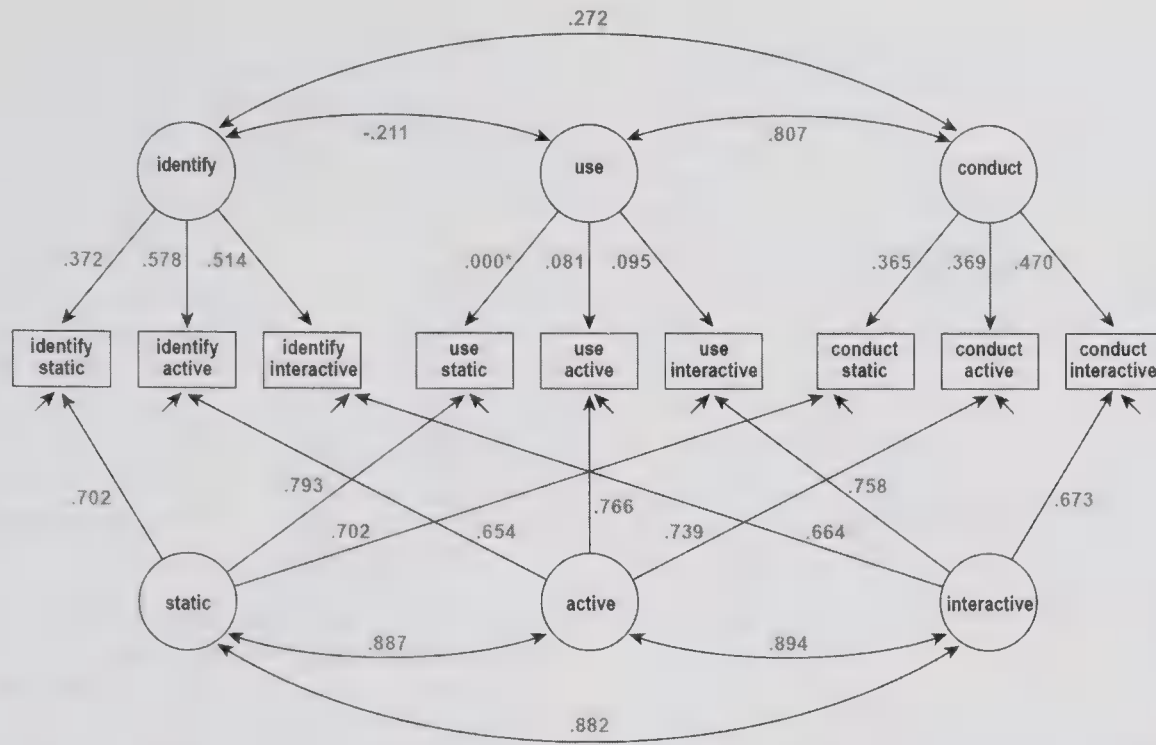


Figure 7. Path model for the multitrait-multimethod confirmatory factor analysis.

items in the interactive modality measure the science practice constructs more distinctly than the other two modalities. In particular, looking at the factor loadings in the column for the *Conducting Inquiry* construct in Table 10, the loading for the interactive modality (.673) is considerably lower than for the static (.702) and active (.739) modalities. This is similar to the finding in the G-study that the interactive modality more distinctly measures the *Conducting Inquiry* construct.

Table 11 shows the multitrait-multimethod (MTMM) correlation matrix based on CFA (the factor loadings and intercorrelations among factors are shown in Figure 7). It provides information about how each science practice is tied to each assessment modality. The figures that are in bold represent the within-modality, cross-science practice correlations. These values are similar as revealed in Table 10.

Multidimensional IRT. Table 12 shows a type of reliability coefficient estimated by the Expected A-Posteriori (EAP)/Plausible Values (PV) derived from the MIRT analysis. The EAP/PV reliability coefficient represents how well the persons (students) are separated by the measures of each of the science practice constructs. As seen in the G-study and the MTMM/CFA, the differences among the assessment modalities are relatively small for the *Identifying* and *Using Principles* constructs, but the Inter-

active modality has a higher reliability coefficient (.82) for the *Conducting* construct. Again, this points to the fact that the interactive assessment modality was more reliable in measuring the *Conducting* science construct than the other two assessment modalities.

Discussion and Implications

With the increasing interest in the use of technology to create assessments that measure skills that are hard to assess in traditional static modalities, this study suggests that engaging students in interactive assessments may provide a better estimate of their more complex inquiry practices than active or static formats do. Such interactive modalities are currently not widely used in science assessment, if at all. The study is grounded in research-based principles and literature sources for informing the design of next generation assessments and an empirical study of the affordances of dynamic and interactive modalities for measuring distinct science practice constructs. These outcomes of the *Foundations of 21st Century Science Assessments* project provide guidelines for designing the next generation of science assessments and evidence supporting claims that the affordances of dynamic, interactive,

Table 9
Factor Loadings of Science Practice Factors on the Sets of Items Represented by Three Assessment Modalities

Science practice factors	Items by assessment modality		
	Static	Active	Interactive
Identifying	0.372	0.578	0.514
Using	0.000	0.081	0.095
Conducting	0.365	0.369	0.470

Table 10
Factor Loadings of Assessment Modality Factors on the Sets of Items Represented by Three Science Practices

Assessment modality factors	Items by science practices		
	Identifying	Using	Conducting
Static	0.702	0.793	0.702
Active	0.654	0.766	0.739
Interactive	0.664	0.758	0.673

Table 11
Multitrait–Multimethod Correlation Matrix Based on CFA

Science practice	Modality	Identifying			Using			Conducting		
		Static	Active	Interactive	Static	Active	Interactive	Static	Active	Interactive
Identifying	Static	1.00								
	Active	0.62	1.00							
	Interactive	0.60	0.69	1.00						
Using	Static	0.56	0.00	0.00	1.00					
	Active	−0.01	0.49	−0.01	0.54	1.00				
	Interactive	−0.01	−0.01	0.49	0.53	0.53	1.00			
Conducting	Static	0.53	0.06	0.05	0.56	0.02	0.03	1.00		
	Active	0.04	0.54	0.05	0.00	0.59	0.03	0.60	1.00	
	Interactive	0.05	0.07	0.51	0.00	0.03	0.55	0.59	0.62	1.00

Note. The figures that are in bold represent the within-modality, cross-science practice correlations. CFA = confirmatory factor analysis.

complex assessment tasks can improve the measurement of science inquiry practices.

We note that the design principles presented in this article relate to the types of summative assessment in the three modalities compared in this study. Literature on the affordances of dynamic, interactive modalities for formative and adaptive purposes was synthesized by the project and will be reported elsewhere. Relatively little research has studied the interaction of multiple media such as text, graphics, and static and dynamic perceptual cuing in complex tasks. There is considerable research to be done on the functions of multiple representations and interactive interfaces in learning and assessments of science systems and practices (Buckley & Quellmalz, 2013).

This study provides rare large-scale evidence that interactive assessments may be more effective than static assessments at discriminating student proficiencies across different types of science practices. Studies comparing item formats have primarily been within the static modality (selected vs. constructed responses) or between performance assessments and conventional tests. This study extends the comparison of task and item design to complex tasks involving inquiry practice constructs and the dynamic and interactive affordances of technology-based complex science assessment tasks. The three modality versions compared (static, active, and interactive) were carefully constructed to keep the representations of the science ecosystem parallel. Thus, all three versions depicted the ecosystems with parallel stylistic images of the organisms, tables, graphs, and screen layouts. In contrast, typical ecosystem items in extant tests tend to vary in the representation of the ecosystem, for example, by presenting a food web as a set of boxes, text organism names, or pictures of organisms. This study aimed for comparable representations of the ecosystems

so that the research variables would be the extent of learner control (static, active, interactive) and the dynamic level of the ecosystem presentation, that is, a still image, an animation, or a dynamic, changing display. Research on design variations within next generation assessments will face similar methodological challenges.

The study of alternative complex task and item formats also presents analysis challenges. In this study, a combination of methods—a generalizability study, MIRT, and confirmatory factor analyses—examined the measurement properties of the modalities through different lenses. Examining the convergent, discriminate, and construct validity of complex, dynamic assessments poses challenges for the measurement community.

Conclusions

This project integrated research on learning in rich multimedia environments with evidence-centered assessment design methods to shape a framework for developing and establishing the technical quality of reusable task designs for assessing complex science learning. We believe this will make a significant contribution to the field by moving the state of technology-based item development to more principled practice through identifying relevant findings from research on model-based reasoning and multimedia learning that affect the design of assessments of learning, retrieval, and transfer.

Current science tests do not address some of the valued knowledge and practices called for in the new *Framework for K-12 Science Education* and draft *Next Generation Science Standards*. Therefore, the next generation of science assessments will need to address both a broader range of standards and innovative methods for assessing them.

The study provides much needed empirical evidence of the affordances of dynamic and interactive assessments for discriminating among science knowledge and inquiry skills. The results suggest that static assessments are not as effective as interactive assessments for differentiating between factual knowledge and the ability to apply that knowledge in meaningful contexts. Our study found that the interactive task sets that served as a basis of the interactive assessments were more effective than either static or active assessments at uniquely measuring students’ ability to engage in inquiry practices. Therefore, assessment developers who wish to design assessments of science inquiry skills should consider the use of active and interactive assessment tasks.

Table 12
EAP/PV Reliability Coefficients for Static, Active, and Interactive Modalities

Modality	Identifying	Using	Conducting
Static	.76	.79	.77
Active	.74	.76	.77
Interactive	.74	.77	.82

Note. EAP/PV = expected a posteriori/plausible values.

References

- Achieve. (2012). *Next generation science standards*. Retrieved from <http://www.nextgenscience.org/next-generation-science-standards>
- Adams, W. K., Reid, S., Lemaster, R., McKagan, S. B., Perkins, K. K., Dubson, M., & Wieman, C. E. (2008). A study of educational simulations: Part I—Engagement and learning. *Journal of Interactive Learning Research*, 19, 397–419.
- Ainsworth, S. (2008). The educational value of multiple-representations when learning complex scientific concepts. In J. K. Gilbert, M. Reiner, & M. Nakhleh (Eds.), *Visualization: Theory and practice in science education* (pp. 191–208). New York, NY: Springer. doi:10.1007/978-1-4020-5267-5_9
- American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.
- Betrancourt, M. (2005). The animation and interactivity principles. In R. E. Mayer (Ed.), *Handbook on multimedia learning* (pp. 287–296). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511816819.019
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.
- Brennan, R. (2001). mGENOVA [Computer software]. Retrieved from <http://www.education.uiowa.edu/centers/casma/computer-programs.aspx>
- Buckley, B. C., Gobert, J., Horwitz, P., & O'Dwyer, L. (2010). Looking inside the black box: Assessing model-based learning and inquiry in BioLogica. *International Journal of Learning Technology*, 5, 166–190. doi:10.1504/IJLT.2010.034548
- Buckley, B. C., & Quellmalz, E. S. (2013). Supporting and assessing complex biology learning with computer-based simulations and representations. In D. Treagust & C.-Y. Tsui (Eds.), *Multiple representations in biological education* (pp. 247–267). Dordrecht, the Netherlands: Springer. doi:10.1007/978-94-007-4192-8_14
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105. doi:10.1037/h0046016
- Clark, R. C., & Mayer, R. E. (2011). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. San Francisco, CA: Pfeiffer. doi:10.1002/9781118255971
- Clement, J. (1989). Learning via model construction and criticism: Protocol evidence on sources of creativity in science. In J. A. Glover, R. R. Ronning, & C. R. Reynolds (Eds.), *Handbook of creativity: Assessment, theory and research* (pp. 341–381). New York, NY: Plenum Press.
- College Board. (2009). *Science: College Boards standards for college success*. Retrieved from <http://professionals.collegeboard.com/profdownload/cbscs-science-standards-2009.pdf>
- Collins, A., Brown, J. S., & Newman, S. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Erlbaum.
- Darling-Hammond, L. (2010). *Performance counts: Assessment systems that support high-quality learning*. Washington, DC: CCSSO.
- Darling-Hammond, L., & Pechione, R. (2010). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. Presented at the National Conference on Next Generation K–12 Assessment Systems, Center for K–12 Assessment & Performance Management, with the Education Commission of the States (ECS) and the Council of Great City Schools (CGCS), Washington, DC.
- Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, 323, 66–69. doi:10.1126/science.1167311
- de Jong, T. (2006). Technological advances in inquiry learning. *Science*, 312, 532–533. doi:10.1126/science.1127750
- Geier, R., Blumenfeld, P., Marx, R., Krajcik, J., Fishman, B., & Soloway, E. (2008). Standardized test outcomes of urban students participating in standards and project based science curricula. *Journal of Research in Science Teaching*, 45, 922–939. doi:10.1002/tea.20248
- Hegarty, M. (2004). Dynamic visualizations and learning: Getting to the difficult questions. *Learning and Instruction*, 14, 343–351. doi:10.1016/j.learninstruc.2004.06.007
- Hmelo-Silver, C. E., Jordan, R., Liu, L., Gray, S., Demeter, M., Rugaber, S., . . . Goel, A. (2008). Focusing on function: Thinking below the surface of complex science systems. *Science Scope*, 31, 27–35.
- Horwitz, P., Gobert, J. D., Buckley, B. C., & O'Dwyer, L. M. (2010). Learning genetics with dragons: From computer-based manipulatives to hypermodels. In M. J. Jacobson & P. Reimann (Eds.), *Designs for learning environments of the future: International perspectives from the learning sciences* (pp. 61–87). New York, NY: Springer. doi:10.1007/978-0-387-88279-6_3
- Ioannidou, A., Repenning, A., Webb, D., Keyser, D., Luhn, L., & Daetwyler, C. (2010). Mr. Vetro: A collective simulation for teaching health science. *International Journal of Computer-Supported Collaborative Learning*, 5, 141–166. doi:10.1007/s11412-010-9082-8
- Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity*, 6, 41–49. doi:10.1002/cplx.1027
- King, K. (2011). *Balanced, multilevel science assessment systems*. Presented at the National Conference on Student Assessment, Orlando, FL.
- Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., . . . Ryan, M. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting learning by design into practice. *The Journal of the Learning Sciences*, 12, 495–547. doi:10.1207/S15327809JLS1204_2
- Koomen, M. (2006). *The development and implementation of a computer-based assessment of science literacy in PISA 2006*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Krajcik, J., Marx, R., Blumenfeld, P., Soloway, E., & Fishman, B. (2000). *Inquiry-based science supported by technology: Achievement and motivation among urban middle school students*. Paper presented at the American Educational Research Association, New Orleans, LA.
- Kriz, S., & Hegarty, M. (2007). Top-down and bottom-up influences on learning from animations. *International Journal of Human-Computer Studies*, 65, 911–930. doi:10.1016/j.ijhcs.2007.06.005
- Kühl, T., Scheiter, K., Gerjets, P., & Edelman, J. (2011). The influence of text modality on learning with static and dynamic visualizations. *Computers in Human Behavior*, 27, 29–35. doi:10.1016/j.chb.2010.05.008
- Larkin, J., & Simon, H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65–100. doi:10.1111/j.1551-6708.1987.tb00863.x
- Lee, H., Plass, J. L., & Homer, B. D. (2006). Optimizing cognitive load for learning from computer-based science simulations. *Journal of Educational Psychology*, 98, 902–913. doi:10.1037/0022-0663.98.4.902
- Lee, O. (2002). Science inquiry for elementary students from diverse backgrounds. In W. G. Secada (Ed.), *Review of research in education* (Vol. 26, pp. 23–69). Washington, DC: American Educational Research Association.
- Lehrer, R., & Schauble, L. (2002). Symbolic communication in mathematics and science: Co-constituting inscription and thought. In E. D. A. J. Byrnes (Ed.), *Language, literacy, and cognitive development. The development and consequences of symbolic communication* (pp. 167–192). Mahwah, NJ: Erlbaum.
- Lehrer, R., Schauble, L., Strom, D., & Pligge, M. (2001). Similarity of form and substance: Modeling material kind. In D. K. S. Carver (Ed.), *Cognition and instruction: 25 years of progress* (pp. 39–74). Mahwah, NJ: Erlbaum.
- Linn, M. C., & Eylon, B.-S. (2011). *Science learning and instruction: Taking advantage of technology to promote knowledge integration*. New York, NY: Routledge.

- Liu, O. L., Lee, H.-S., & Linn, M. C. (2011). Measuring knowledge integration: Validation of four-year assessments. *Journal of Research in Science Teaching*, 48, 1079–1107. doi:10.1002/tea.20441
- Loehlin, J. C. (1998). *Latent variable model: An introduction to factor, path, and structural analysis*. Mahwah, NJ: Erlbaum.
- Lowe, R., & Schnotz, W. (2008). *Learning with animation: Research implications for design*. Cambridge, NY: Cambridge University Press.
- Lowe, R. K., & Schnotz, W. (Eds.). (2007). *Learning with animation*. New York, NY: Cambridge University Press.
- Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., Geier, R., & Tal, R. T. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of Research in Science Teaching*, 41, 1063–1080. doi:10.1002/tea.20039
- Mayer, R. E. (Ed.). (2005a). *The Cambridge handbook of multimedia learning*. New York, NY: Cambridge University Press. doi:10.1017/CBO9780511816819
- Mayer, R. E. (2005b). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511816819.004
- Mayer, R. E., & Johnson, C. (2008). Revising the redundancy principle in multimedia learning. *Journal of Educational Psychology*, 100, 380–386. doi:10.1037/0022-0663.100.2.380
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 12–23.
- Metz, K. E. (2004). Children's understanding of scientific inquiry: Their conceptualization of uncertainty in investigations of their own design. *Cognition and Instruction*, 22, 219–290. doi:10.1207/s1532690xci2202_3
- Minnesota Department of Education. (2010). *Draft test specifications for science*. Retrieved from http://education.state.mn.us/MDE/Accountability_Programs/Assessment_and_Testing/Assessments/MCA/TestSpecs/index.html
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design*. Princeton, NJ: Educational Testing Service.
- National Assessment Governing Board. (2008). *Science framework for the 2009 national assessment of educational progress*. Washington, DC: National Assessment Governing Board.
- National Center for Education Statistics. (2012). *The nation's report card: Science in action: Hands-on and interactive computer tasks from the 2009 science assessment* (Report No. NCES 2012–468). Washington, DC: Institute of Education Sciences, U.S. Department of Education.
- National Research Council. (2011). *A framework for k-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.
- Nersessian, N. J. (2008). *Creating scientific concepts*. Cambridge, MA: MIT Press.
- Papaevripidou, M., Constantinou, C. P., & Zacharia, Z. C. (2007). Modeling complex marine ecosystems: An investigation of two teaching approaches with fifth graders. *Journal of Computer Assisted Learning*, 23, 145–157. doi:10.1111/j.1365-2729.2006.00217.x
- Pedone, R., Hummel, J. E., & Holyoak, K. J. (2001). The use of diagrams in analogical problem solving. *Memory & Cognition*, 29, 214–221. doi:10.3758/BF03194915
- Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Plass, J. L., Homer, B. D., & Hayward, E. O. (2009). Design factors for educationally effective animations and simulations. *Journal of Computing in Higher Education*, 21, 31–61. doi:10.1007/s12528-009-9011-x
- Quellmalz, E. S., & Haertel, G. (2004). *Technology supports for state science assessment systems*. Unpublished manuscript.
- Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science*, 323, 75–79. doi:10.1126/science.1168046
- Quellmalz, E., Schank, P., Hinojosa, T., & Padilla, C. (1999). *Performance assessment links in science* (PALS; Report No. EDO-TM-99–04). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.
- Quellmalz, E. S., Timms, M. J., & Buckley, B. (2010). The promise of simulation-based science assessment: The calipers project. *International Journal of Learning Technology*, 5, 243–263. doi:10.1504/IJLT.2010.037306
- Quellmalz, E. S., Timms, M. J., Buckley, B. C., Davenport, J., Loveland, M., & Silberglitt, M. D. (2011). 21st century dynamic assessment. In J. Clarke-Midura, M. Mayrath, & C. Dede (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 55–89). Charlotte, NC: Information Age.
- Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching*, 49, 363–393.
- Rieber, L. P., Tzeng, S., & Tribble, K. (2004). Discovery learning, representation, and explanation within a computer-based simulation. *Computers and Education*, 27, 45–58.
- Rivet, A., & Krajcik, J. S. (2004). Achieving standards in urban systemic reform: An example of a sixth grade project-based science curriculum. *Journal of Research in Science Teaching*, 41, 669–692. doi:10.1002/tea.20021
- Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 116–136. doi:10.1037/0278-7393.25.1.116
- Schwartz, D. L., & Heiser, J. (2006). Spatial representations and imagery in learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 283–298). Cambridge, England: Cambridge University Press.
- Simon, H. A. (1980). Problem solving and education. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 81–96). Hillsdale, NJ: Erlbaum.
- Slotta, J. D., & Chi, M. T. H. (2006). Helping students understand challenging topics in science through ontology training. *Cognition and Instruction*, 24, 261–289. doi:10.1207/s1532690xci2402_3
- Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives*, 4, 1–98. doi:10.1080/15366367.2006.9678570
- Stewart, J., Cartier, J. L., & Passmore, C. M. (2005). Developing understanding through model-based inquiry. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn* (pp. 515–565). Washington, DC: The National Academies Press.
- Tversky, B., Heiser, J., Lozano, S., MacKenzie, R., & Morrison, J. (2008). Enriching animations. In R. K. Lowe & W. Schnotz (Eds.), *Learning with animation: Research and design implications* (pp. 263–285). New York, NY: Cambridge University Press.
- van Merriënboer, J. J. G., & Kester, L. (2005). The four-component instructional design model: Multimedia principles in environments for complex learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 71–94). New York, NY: Cambridge University Press. doi:10.1017/CBO9780511816819.006
- Vattam, S. S., Goel, A. K., Rugaber, S., Hmelo-Silver, C. E., Jordan, R., Gray, S., & Sinah, S. (2011). Understanding complex natural systems by articulating structure-behavior-function models. *Journal of Educational Technology & Society*, 14, 66–81.
- Zacharia, Z. C. (2007). Comparing and combining real and virtual experimentation: An effort to enhance students' conceptual understanding of electric circuits. *Journal of Computer Assisted Learning*, 23, 120–132. doi:10.1111/j.1365-2729.2006.00215.x

Received December 15, 2011

Revision received January 10, 2013

Accepted February 12, 2013 ■

My Science Tutor: A Conversational Multimedia Virtual Tutor

Wayne Ward

Boulder Language Technologies, Boulder, Colorado, and
University of Colorado at Boulder

Ron Cole, Daniel Bolaños,

Cindy Buchenroth-Martin, and Edward Svirsky
Boulder Language Technologies

Tim Weston

University of Colorado at Boulder

My Science Tutor (MyST) is an intelligent tutoring system designed to improve science learning by elementary school students through conversational dialogs with a virtual science tutor in an interactive multimedia environment. Marni, a lifelike 3-D character, engages individual students in spoken dialogs following classroom investigations using the kit-based Full Option Science System program. MyST attempts to elicit self-expression from students; process their spoken explanations to assess understanding; and scaffold learning by asking open-ended questions accompanied by illustrations, animations, or interactive simulations related to the science concepts being learned. MyST uses automatic speech recognition, natural language processing, and dialog-modeling technologies to interpret student responses and manage the dialog. Sixteen 20-min tutorials were developed for each of 4 areas of science taught in 3rd, 4th, and 5th grades. During summative evaluation of the program, students received one-on-one tutoring via MyST or an expert human tutor following classroom instruction on the science topic, representing over 4.5 hr of tutoring across the 16 sessions. A quasi-experimental design was used to compare average learning gain for 3 groups: human tutoring, virtual tutoring, and no tutoring. Learning gain was measured using standardized assessments given to students in each condition before and after each science module. Results showed that students in both the human and virtual tutoring groups had significant learning gains relative to students in the control classrooms and that there were no significant differences in learning gains between students in the human and MyST human tutoring conditions. Both teachers and students gave high-positive survey ratings to MyST.

Keywords: intelligent tutors, spoken dialog, science learning

According to the 2009 National Assessment of Educational Progress (NAEP, 2005), only 34% of fourth graders, 30% of eighth graders, and 21% of 12 graders tested as proficient in

science, with 1%–2% of these students demonstrating advanced knowledge of science in these grades. Thus, over two thirds of U.S. students are not proficient in science. The vast majority of these students are in low-performing schools that include a high percentage of disadvantaged students from families with low socioeconomic status, which often include English learners with low English-language proficiency. Analysis of the NAEP scores in reading, math, and science over the past 20 years indicate that this situation is getting worse. For example, the gap between English learners and English-only students, which is over one standard deviation lower for English learners, has increased rather than decreased over the past 20 years. Moreover, science instruction is often underemphasized in U.S. schools, with reading and math being stressed. My Science Tutor (MyST) was designed to address this problem by immersing students in a multimedia environment with a virtual science tutor that was designed to behave like an engaging and effective human tutor. The focus of the program is to improve each student's engagement, motivation, and learning by helping them learn to visualize, reason about, and explain science during conversations with the virtual tutor.

The learning principles embedded in MyST are consistent with conclusions and recommendations of the National Research Council Report, "Taking Science to School: Learning and Teaching Science in Grades K-8" (Duschl, Schweingruber, & Shouse,

This article was published Online First September 9, 2013.

Wayne Ward, Boulder Language Technologies, Boulder, Colorado, and Computational Language and Education Research Center, University of Colorado at Boulder; Ron Cole, Daniel Bolaños, Cindy Buchenroth-Martin, and Edward Svirsky, Boulder Language Technologies; Tim Weston, ATLAS Center, University of Colorado at Boulder.

The research reported here was supported by Institute of Education Sciences, U.S. Department of Education Award R305B070008, National Science Foundation Grants DRL 0733323, awarded to Boulder Language Technologies, and DRL 0733322, awarded to the University of Colorado at Boulder, and NIH Award R44 HD055028 to Mentor Interactive Inc. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute of Education Sciences or the National Science Foundation. We gratefully acknowledge the help of Angel Stobaugh, Director of Literacy Education at Boulder Valley School District, and the principals and teachers who allowed us to visit their schools and classrooms. We appreciate the amazing efforts of Jennifer Borum, Linda Hill, Suzan Heglin, and the rest of the human experts who scored the text passages.

Correspondence concerning this article should be addressed to Wayne Ward, Boulder Language Technologies, 2960 Center Green Court, Boulder, CO 80301. E-mail: wward@bltek.com

2007), which emphasizes the critical importance of scientific discourse in K–12 science education. The report identifies the following crucial principles of scientific proficiency:

Students who are proficient in science: 1. know, use, and interpret scientific explanations of the natural world; 2. generate and evaluate scientific evidence and explanations; 3. understand the nature and development of scientific knowledge; and 4. participate productively in scientific practices and discourse. (p. 2)

The report also emphasizes that *scientific inquiry and discourse is a learned skill*, so students need to be involved in activities in which they learn appropriate norms and language for productive participation in scientific discourse and argumentation.

In a meta-analysis of 18 studies by Chi (2009), the author examined student learning along the continuum *active, constructive, interactive*. Active tasks include “doing something,” such as participating in a classroom science investigation. Constructive tasks include “producing something,” such as a written report describing the results of the investigation. Interactive tasks require discourse and argumentation with a peer or tutor. Chi’s analysis of the research studies produced strong evidence that interactive tasks produce the greatest learning gains.

A substantial body of research indicates that engaging in discourse and argumentation about science is one of the most challenging tasks for young learners, and one of the most important and beneficial skills for them to acquire (Hake, 1998; Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009; Osborne, 2010; Soter et al., 2008). However, evidence also indicates that authentic conversations are extremely rare across all content areas in U.S. classrooms (Cazden, 1988; Gamoran & Nystrand, 1991; Nystrand, 1997). As Osborne (2010) noted, “Argument and debate are common in science, yet they are virtually absent in science education” (p. 463). Our goal in designing MyST was to provide students with the scaffolding, modeling, and practice they need to learn to reason and talk about science.

MyST is an intelligent tutoring system intended to provide an intervention for third-, fourth-, and fifth-grade children who are struggling with science. In our study, it was used as a supplement to normal classroom instruction using the Full Option Science System (FOSS). FOSS is an inquiry-based science program that is based on the idea that “The best way for students to appreciate the scientific enterprise, learn important scientific concepts, and develop the ability to think well is to actively construct ideas through their own inquiries, investigations and analyses” (FOSS, n.d., para. 3). It has been under development since 1988, and is in use in every state in the United States. Twenty-six science modules have been developed for Grades K–6. The learning objectives in each FOSS module are aligned to the National Science Education Standards and standards for most states. Each module covers an integrated area of science (e.g., Mixtures and Solutions, Measurement, Variables). The instructional materials for each module are packaged in a kit that contains the materials needed to conduct the classroom science investigations: a teacher guide, a module-specific teacher-preparation video, and a summative assessment (Assessing Science Knowledge [ASK]) to be administered before and after each science module.

Within a science module, students in classrooms work in small groups to conduct a series of approximately 16 science investigations over an 8- to 10-week period. These hands-on investigations

are aligned to specific science concepts and learning objectives. The structure of the FOSS program provides an ideal test bed for research and evaluation of MyST, with MyST dialogs being aligned with specific classroom science investigations, learning objectives, science standards, and ASK assessments.

Research Motivating the Design of MyST Dialogs

MyST is an example of a new generation of intelligent tutoring systems that facilitate learning through natural spoken dialogs with a virtual tutor in multimedia activities. Intelligent tutoring systems aim to enhance learning achievement by providing students with individualized and adaptive instruction similar to that provided by a knowledgeable human tutor. These systems support typed or spoken input, with the system presenting prompts and feedback via text, a human voice, or an animated pedagogical agent (Graesser, VanLehn, Rosé, Jordan, & Harter, 2001; Lester et al., 1997; Mostow & Aist, 2001; VanLehn et al., 2007; Wise et al., 2005). Text, illustrations, and animations may be incorporated into the dialogs. Research studies show up to one sigma gains (approximately equivalent to an improvement of one letter grade) when comparing performance of high school and college students who use the tutoring systems with students who receive classroom instruction on the same content (Graesser et al., 2001; VanLehn & Graesser, 2001; VanLehn et al., 2005). In a recent synthesis of research that compared learning gains following human tutoring or following use of an intelligent tutoring system, VanLehn (2011) concluded that human tutoring and intelligent tutoring systems produce approximately the same effect size, with human tutoring at $d = 0.79$ and intelligent tutoring systems at $d = 0.76$.

The development of MyST is informed by several decades of research in psychology and computer science. In the remainder of this section, we briefly describe theory and research that informed the design of MyST.

Benefits of Tutorial Instruction

Theory and research provide strong guidelines for designing effective tutoring dialogs. Over two decades of research have demonstrated that learning is most effective when students receive individualized instruction in small groups or one-on-one tutoring. Bloom (1984) determined that the difference between the amount and quality of learning for students who received classroom instruction and those who received either one-on-one or small-group tutoring was two standard deviations. Evidence that tutoring works has been obtained from dozens of well-designed research studies, meta-analyses of research studies (Cohen, Kulik, & Kulik, 1982), and positive outcomes obtained in large-scale tutoring programs (Madden & Slavin, 1989; Topping & Whiteley, 1990).

Benefits of tutoring can be attributed to several factors, including the following:

Question generation. A significant body of research shows that learning improves when teachers and students ask deep-level-reasoning questions (Bloom, 1956). Asking authentic questions leads to improved comprehension, learning, and retention of texts and lectures by college students (Craig, Gholson, Ventura, & Graesser, 2000; Driscoll et al., 2003; King, 1989) and school children (King, 1994; King, Staffieri, & Adelgais, 1998; Palinscar & Brown, 1984).

Generating explanations. Research has demonstrated that having students produce explanations improves learning (Chi et al., 1989; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; King, 1994; King et al., 1998; Palinscar & Brown, 1984). In a series of studies, Chi et al. (1989, 2001) found that having college students generate self-explanations of their understanding of physics problems improved learning. Self-explanation also improved learning about the circulatory system by eighth-grade students in a controlled experiment (Chi, De Leeuw, Chiu, & LaVancher, 1994). Hausmann and Van Lehn (2007a, 2007b) note that “self-explaining has consistently been shown to be effective in producing robust learning gains in the laboratory and in the classroom” (2007b, p. 1067.) Experiments by Hausmann and Van Lehn (2007b) indicate that it is the process of actively producing explanations, rather than the accuracy of the explanations, that makes the biggest contribution to learning.

Knowledge coconstruction. Students coconstruct knowledge when they are provided with the opportunity to express their ideas and to evaluate their thoughts in terms of ideas presented by others. There is compelling evidence that engaging students in meaningful conversations improves learning (Butcher, 2006; Chi et al., 1989; King, 1994; King et al., 1998; Murphy et al., 2009; Palinscar & Brown, 1984; Pine & Messer, 2000; Soter et al., 2008).

Social Constructivism

In social constructivism, learning is viewed as an active social process of constructing knowledge “that occurs through processes of interaction, negotiation, and collaboration” (Palinscar, 1998, p. 365). Vygotsky (1978) stressed the critical role of social interaction within one’s culture in acquiring the social and linguistic tools that are the basis of knowledge acquisition. “Learning awakens a variety of internal developmental processes that are able to operate only when the child is interacting with people in his environment” (Vygotsky, 1978, pp. 89–90). He stressed the importance of having students learn by presenting problems that enable them to scaffold existing knowledge to acquire new knowledge. Vygotsky introduced the concept of the zone of proximal development, “the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers” (Vygotsky, 1978, p. 86). Social constructivism provides the conceptual model for knowledge acquisition in MyST: to improve learning by scaffolding conversations using open-ended questions and media to support hypothesis generation and coconstruction of knowledge.

Discourse Comprehension Theory

Cognitive learning theorists generally agree that learning occurs most effectively when students are actively engaged in critical thinking and reasoning processes that cause new information to be integrated with prior knowledge. Discourse comprehension theory (Kintsch, 1988, 1998) holds that deep learning requires integration of prior knowledge with new information and results in the ability to use this information constructively in new contexts. To the extent possible, MyST attempts to determine relevant information that students know and build on that lead students to correct explanations.

Social Agency and Pedagogical Agents

When human computer interfaces are consistent with the social conventions that guide our daily interactions with other people, they provide more engaging, satisfying, and effective user experiences (Nass & Brave, 2005; Reeves & Nass, 1996). Such programs foster social agency, enabling users to interact with them the way they interact with people. In comparisons of programs with and without talking heads or human voices, children learned more and reported more satisfaction using programs that incorporated virtual humans (Atkinson, 2002; Baylor & Kim, 2005; Moreno, Mayer, Spire, & Lester, 2001). A number of researchers have observed that children become highly engaged with virtual tutors and appear to interact with a virtual tutor as if it were a real teacher and appear motivated to work hard to please it. Lester (Lester et al., 1997) termed this phenomenon the “persona effect.”

Multimedia Learning

During MyST dialogs, students are encouraged to construct explanations of science presented in illustrations, silent animations, and interactive simulations. The design of these dialogs is consistent with research indicating that combining spoken explanations with media can optimize science learning, either during multimedia presentations (Horz & Schnotz, 2010; Mayer, 2001, 2005) or when students are required to generate explanations in multimedia learning environments (Roy & Chi, in press). In a series of studies, Mayer (2001) investigated students’ ability to learn how things work (motors, brakes, pumps, lightning) when information was presented in different modalities (e.g., text with illustrations, or narration of the text during which a spoken voice explained the information presented in an illustration or sequence of illustrations). A key finding of Mayer’s work is that simultaneously presenting speech (narration) with nonverbal visual information (a sequence of illustrations or an animation) results in the highest retention of information and the application of knowledge to new problems. Mayer (2001) argued that when a person is presented with a well-designed narrated animation, the listener is able to construct an enriched multimodal representation of the two sources of input, leading to superior recall and transfer of knowledge to new tasks. Roy and Chi (in press), based on a review of the literature on self-explanations in multimedia environments, suggest that

many learners would benefit from self-explanation training or prompting within multimedia environments. Essentially, we have argued that because they are information rich, multimedia environments afford the generation of many opportunities for explaining encoded information and accessing and relating prior knowledge. (p. 27)

Dialog Interaction

The design of spoken dialogs in MyST is based on a number of principles used in Questioning the Author (QtA), an approach to classroom discussions developed by Isabel Beck and Margaret McKeown (Beck, McKeown, Sandora, Kucan, & Worthy, 1996; McKeown & Beck, 1999; McKeown, Beck, Hamilton, & Kucan, 1999). During the 3-year period in which MyST dialogs were designed, tested, and refined, we worked with QtA codeveloper Margaret McKeown to apply principles of QtA to spoken dialogs

with Marni that incorporate illustrations, animations, and interactive simulations to help students visualize the science they are trying to explain.

QtA is a mature, scientifically based, and effective program used by hundreds of teachers across the United States. It is designed to improve comprehension of narrative or expository texts that are discussed as they are read aloud in the classroom. The focus is to have students grapple with, and reflect on, what an author is trying to say in order to build a representation from the text. The approach uses open-ended questions to initiate discussion (What is the author trying to say?) to help students focus on the author's message (That's what she says, but what does she mean?) to help students link information (How does that fit with what the author already told us?) and to help the teacher guide students toward comprehension of the text.

QtA provides a good basis for tutorial interaction in the MyST virtual tutoring system because (a) research shows that it is effective for improving comprehension (Murphy & Edwards, 2005); (b) it provides a framework and planning process that helps define learning goals and develops an orderly sequence for getting students to achieve the goals; (c) it offers ways to design prompts that draw student attention to relevant portions of presented material, but that are open enough to leave the identification of the material to students; (d) it provides a principled, easily understandable and well-documented program for teachers or tutors to elicit and respond to student responses that helps them learn to focus on and make connections between meaningful elements of the discourse and their own experiences; and (e) it focuses on comprehension, with discussion of student personal views and experiences limited to those that can directly enhance building meaning from texts, lectures, multimedia presentations, data sets, or hands-on learning activities.

Murphy and Edwards (2005) analyzed the results of research studies that met rigorous scientific criteria for evaluating programs designed to improve student learning through classroom conversations. Of the nine programs that met the scientific criteria for valid research studies, QtA was identified as one of two approaches that is likely to promote high-level thinking and comprehension of text (Murphy & Edwards, 2005). Moreover, analysis of the QtA discourse showed a relatively high incidence of authentic questions, uptake, and teacher questions that promoted high-level thinking—all indicators of productive discussions likely to promote learning and comprehension of text (Soter & Rudge, 2005).

The MyST System

System Description

Students learn science in MyST through natural spoken dialogs with the virtual tutor Marni, a 3-D computer character that is on screen at all times. Marni asks students open-ended questions related to illustrations, silent animations, or interactive simulations displayed on the computer screen. Figure 1 displays a screen shot of Marni asking questions about media displayed in a tutorial. The student's computer shows a full screen window that contains Marni, a display area for presenting media, and a display button that indicates the listening status of the system. Marni produces accurate visual speech, with head and face movements that are synchronized with her speech.

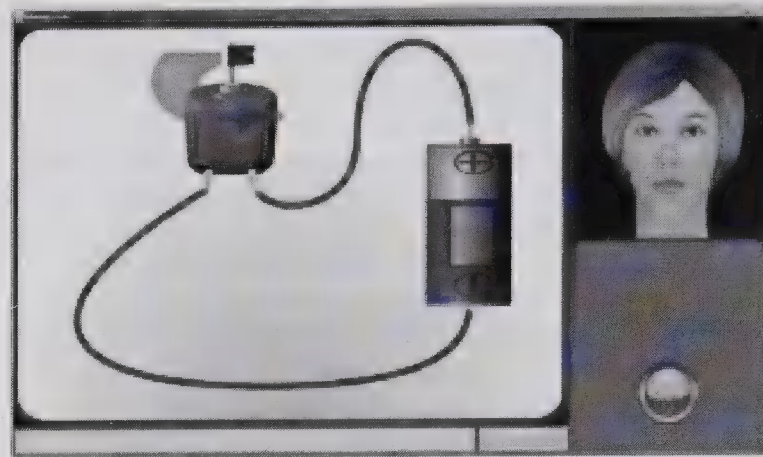


Figure 1. My Science Tutor (MyST) screen layout.

We call these conversations with Marni *multimedia dialogs*, because students simultaneously listen to and think about Marni's questions while viewing illustrations and animations or interacting with a simulation. The media facilitate dialogs with Marni by helping students visualize the science they are discussing. The primary focus of each dialog is to elicit self-explanations from students. MyST analyzes the spoken explanations to determine what the student does and does not know about the science, then presents follow-up questions, which may be accompanied by new media, to help the student construct a correct explanation of the phenomena being studied. The virtual tutor Marni, who speaks with a recorded human voice, is designed to behave like an effective human tutor that the student can relate to and work with to learn science. This is achieved by modeling dialogs between students and human tutors trained in using QtA during the development phase of the project. These dialogs scaffold learning by providing students with support when needed until they can apply new skills and knowledge independently (Vygotsky, 1978).

Marni elicits self-explanations from students using strategies that embody QtA dialog moves such as *marking* and *revoicing*. These two techniques require that the system identify the student's dialog content (marking it) followed by repeating (revoicing) a paraphrase of the information back to the student as a part of the next question: *You mentioned that electricity flows in a closed path. What else can you tell me about how electricity flows?* Marni's responses are designed to communicate this understanding back to the students and to engage and assure them that she understands what they are saying.

A tutorial session generally begins with relating the session to what the student has recently covered in class (during a science investigation), with Marni saying something like: *What have you been studying in science recently?* If the student says something recognizable as the tutorial topic (e.g., "We made a circuit"), the system moves forward by asking the student what they know about the topic: *You mentioned circuits. Can you tell me what a circuit is?* If nothing from what the system extracted from the student's answer relates to the topic, then Marni introduces the topic: *I heard you were learning about circuits. Can you tell me what a circuit is?* For each key concept discussed, the interaction typically begins with a general open-ended question (accompanied by media, such as a picture of a simple circuit): *What's this all about?* or *What's going on here?* and then proceeds to more directed open-ended

questions like: Can you tell me more about the flow of electricity in the circuit?

Media are used to ground the conversation, focus the student's attention, help the student visualize the science, and provide a visual frame of reference for the student to talk about. The media are not narrated, and they do not explain the concept to the student. A typical strategy used by MyST is to show an animation to the student and ask him or her to explain what is going on. The use of media was initially intended as a mechanism to get students past *sticking points*, points in a dialog when the system is not able to elicit information from the student that it can build on. During dialogs with project tutors during system development, discussed below, the method proved so useful for eliciting explanations that tutors began to use this as the standard introduction to concepts: ask an introductory question about what a student knows, show an illustration, and ask what is going on.

As noted, MyST dialogs incorporate three types of media: (a) illustrations, (b) animations, and (c) interactive simulations, illustrated in Figure 2. Although these sometimes overlap in the content presented, each plays a unique role. Illustrations are static Flash drawings and are a good way to initiate discussions about topics. They provide the student with a visual frame of reference that helps focus the student's attention and the subsequent discussion on the content of the illustration: *So, what's going on here?* Animations are noninteractive, silent Flash animations that help students visualize concepts that can be difficult to capture in illustrations. In Figure 2, the direction of the flow of electricity is represented by blue dots moving from the D-cell through the wires and bulb and back to the D-cell. The animations enable Marni to ask the student questions to elicit explanations about what is being shown. Simulations allow students to interact directly with the Flash animation using a mouse. Figure 2 shows a simulation of a FOSS classroom investigation called "Breaking the Force" in which students investigate how much weight (number of metal washers) is required on one side of a balance scale to break the force of the magnets attracting each other on the other side. The number of washers in the cup and the space between magnets can be investigated and graphed in this simulation. During multimedia dialogs, as students are interacting with a simulation, the tutor can say things like: *What could you do to . . . ? What happens if you . . . ?*

System Operation (How Spoken Dialogs Work)

MyST uses character animation, automatic speech recognition, natural language processing, and dialog modeling to support con-

versations with Marni. The dialogs are designed to elicit responses from students that show their understanding of a specific set of points. The key points of a dialog are specified as propositions realized as semantic frames. The frames represent the events and entities in the domain and the roles that they play. For example, *Current goes from the negative terminal to the positive* would be represented as: **Electricity Flows Origin.negative Destination.positive**. During spoken dialogs, the tutor asks questions that are designed to elicit student responses that will map to the elements of the targeted semantic frames. Information extracted from student responses is integrated into the session context that represents which points have been addressed by the student, which have not, which were expressed correctly, and which represented misconceptions. In analyzing a student's answer, the system tests whether the correct values are filling the semantic roles (i.e., whether the value of Origin is negative or positive). On the basis of the current context, the system generates questions to elicit explanations of the elements needed to produce a complete explanation. Follow-up questions and media presentations are designed to scaffold learning by providing hints about the important elements of the investigation that the student did not include or misunderstood. When possible, the follow-up questions are created by taking a relevant part of the student's response and asking for elaboration, explanation, or connections to other ideas.

This interaction style is well suited to automatic speech recognition (ASR) technology, which will have some amount of recognition error. In sessions in which the system is able to accurately recognize and parse student responses, it is able to adapt the tutorial to the individual student. It may move on to another point or delve more deeply into a discussion of concepts that were not correctly expressed by the student, using marking and revoicing to incorporate information from the student's response. If the student does not seem to grasp the basic elements under discussion, the system presents more background material. If the system is unable to elicit and understand relevant student responses, by default it proceeds through the session with a full discussion of each point.

Using spoken responses in this way can increase efficiency and naturalness of the interaction while minimizing the impact of system errors. False-negative errors, in which the system does not recognize correct information provided by the student, simply cause the system to continue to talk about the same point in a different way rather than moving on. False-accept errors, where the system fills in an element because of a recognition error, may cause the system to move on from a point before it is sufficiently

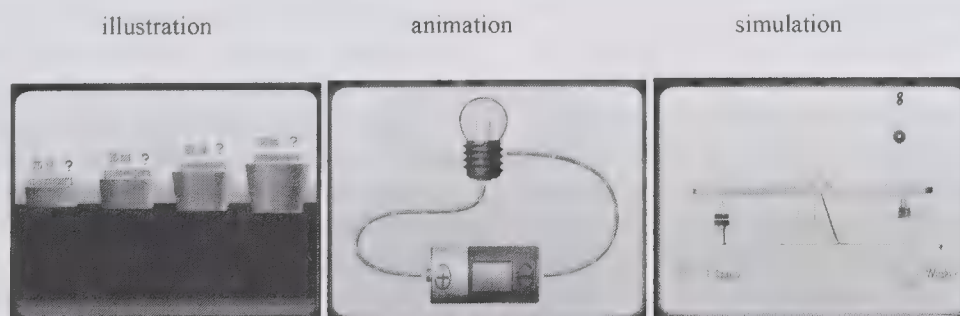


Figure 2. Media types.

covered. False-accept errors are rare and have not proved to be a problem.

System Development

During the development and evaluation of MyST, data were collected from tutoring sessions at elementary schools in the Boulder Valley School District (BVSD). A team of project tutors was trained in the FOSS content and QtA-based interaction style. Using FOSS teacher guides, the team developed learning objectives and specifications for media presentations aligned to each classroom science investigation. Tutors went into the schools and tutored students using the materials developed. Visuals were presented on laptops, and students wore headsets for recording their speech. The recorded sessions were reviewed in group meetings to revise the presentations and determine *sticking points* that would benefit from the introduction of media. These meetings also helped foster a common style across tutors. In addition, transcripts of tutoring sessions were reviewed and annotated by M. McKeown to provide constructive feedback to the project tutors on how to use QtA principles most effectively. The data collected in the human-tutored sessions were used to train the speech recognition and natural language-processing modules to interpret the students' speech and to develop dialog models to attempt to emulate the behavior of the human tutors. These modules were integrated to produce the first version of MyST that was used in Wizard-of-OZ (WOZ) studies.

WOZ

WOZ data collection attempts to provide user interactions similar to the target application, but a human controls the system behavior. In the WOZ collection, students independently interacted with Marni, while a remote human tutor, connected to the student's computer via the Internet, monitored and controlled the system's behavior. The human wizard could see everything on the student's computer and hear what the student was saying. At each point in a dialog when the system was about to take an action (e.g., have Marni talk; present a new illustration), the action was first shown to the human wizard who could accept or change the action. The system logged all transactions during the session. Transcriptions of the dialogs in each session were then reviewed by developers to refine the dialog model. The primary changes during this phase of development included adding new media, expanding the coverage of the natural language processing (to accommodate new ways students could talk about concepts), and adding new ways of asking students questions. As the tutorials evolved, human wizards intervened less.

In sum, during initial development of tutorial dialogs with human tutors, a total of 189 students received human tutoring over a total of 427 sessions. During the subsequent WOZ sessions, a total of 347 students received WOZ tutoring over 1,156 sessions. The purpose of data collected during development was to improve system coverage, that is, modeling the different ways that diverse students talked about science and refine the media presentations, so the emphasis was on including a greater variety of students, with less data from each individual student than in the system evaluation.

System Evaluation

All data collected in the human-tutoring and WOZ sessions were used to train the final acoustic, language, and dialog models for the virtual tutoring system. During the 2010–2011 school year, an assessment of the MyST system was conducted to examine the effect of the virtual tutor on student test scores in science. During the assessment, students interacted with Marni independently in their schools, without a human wizard. An experimenter logged students into the MyST system and specified the dialog session to be used, but otherwise left students alone to use the system. The experimental design compared students receiving MyST tutoring with those receiving face-to-face human tutoring in small groups.

Students were randomly assigned within classrooms to tutoring condition, and these groups were also compared with students from intact control classrooms with no tutoring. Students completed one of four FOSS modules (*Variables, Magnetism, and Electricity, Measurement and Water*) and were tested pre–post with the FOSS-ASK assessment for that module. All students received similar classroom instruction. The two hypotheses for the study were as follows:

Hypothesis 1: Students receiving tutoring with MyST will show learning gains roughly similar to students receiving face-to-face human tutoring.

Hypothesis 2: Both groups receiving tutoring will show greater learning gains than students receiving no tutoring.

Method

Participants

Data were collected from tutoring sessions at elementary schools in the BVSD. BVSD is a 27,000-student school district with 34 elementary schools. There is substantial student diversity across schools, which vary from low to high performing on state science tests. A list of potential schools was developed in collaboration with the BVSD science director. All third-, fourth-, and fifth-grade teachers at these schools were invited to participate in the study, and teachers who accepted were enrolled in the study. All students in the classrooms of participating teachers were invited to participate. All students who agreed to participate were enrolled. All third-, fourth-, and fifth-grade teachers in the district who did not participate as treatment classrooms were recruited to serve as control classrooms, and those who agreed were enrolled.

The data set contained 1,478 students at 22 schools and 63 classrooms. One hundred two students in 14 classrooms in six schools were tutored with MyST, and 85 students in these same classrooms received human tutoring. Control students accounted for 1,155 students in 49 classrooms and 19 schools. These students received no tutoring, but did receive instruction in FOSS modules during class. For analysis, nonconsented students were removed from the sample. Other reasons for removing students from the sample included unmatched pre–post tests where students did not fill out a majority of answers and tests with grading concerns, including very low reliabilities. The remaining sample totaled 1,167 students. Eighty-three students received MyST tutoring, 69 were tutored in small groups

(both in 12 classrooms), and 1,015 students in 50 classrooms in 20 schools received only classroom instruction and no tutoring. All missing data were removed by an analyst who was blind to the experimental condition.

Procedure

Consented students in the study were assigned to receive tutoring *in addition to* their normal classroom instruction for the module. Teachers specified the space in the school to be used, and this varied from school to school, generally any relatively quiet room. The teacher also scheduled the time for their students to minimize the impact on the student's other activities. Tutoring times were always during regular school hours. General guidelines were that this time should not be at recess or lunch, during core subject time (reading, math, science), or during special activities time (art, music).

All students in the study received in-class instruction in the FOSS modules: Measurement (third grade), Magnetism and Electricity (fourth grade), Water (fourth grade), and Variables (fifth grade). Teachers in both treatment and control classrooms followed module lesson plans and used FOSS materials. Students participating in the study received tutoring from MyST or human tutors for 12–16 20-min sessions concurrent with their regular classroom instruction. Each tutorial was oriented around a set of key concepts the student was expected to have learned from classroom instructional activities. Both MyST and human tutoring used the same multimedia content linked to FOSS content. MyST students were tutored individually on computers. Headsets with earphones and microphones were used to reduce noise interference. For most sessions, eight students at a time used the computers in a separate resource room at each school. Students in the human tutoring condition received tutoring with human tutors for the same amount of time as those in the MyST group. They worked in groups of three to four students with each human tutor. Although one-on-one interaction with a human tutor would present a more direct comparison to the virtual tutor condition, the study did not have sufficient resources to provide one-on-one human tutoring; however, research has demonstrated equivalent learning gains for one-on-one and small-group tutoring (e.g., Bloom, 1984).

Measures

Students in all experimental groups were given the ASK summative assessments as pre- and posttest measures. Tests were administered before the beginning of the FOSS lessons for the module, and immediately after tutoring for the module ended. The ASK assessments for the four modules used in the assessment have identical pre and post versions. Depending on the module, the assessments have between eight and 12 items, consisting of multiple-choice and constructed response questions, and show composite internal reliability with alphas in the range of 0.80–0.90. The interrater reliability for subjective items has also met high standards in similar conditions (e.g., $r = .90$), and the validity of the measures has been built up over time through a process of empirical investigation.

Because module tests have different scales, scores were standardized to a common metric. All standardization was conducted

on data with outliers and other spurious data removed. “Testwise” standardization subtracted the mean of each test (over all students and pooling pre/post) from each student's score. This difference was then divided by the average standard deviation for both pre and post for each test.

Pairs of raters (tutors) scored all assessments from tutored students and a subset of assessments from control students. Raters trained together with scoring rubrics provided by FOSS, then scored the assessments independently. All scoring was blind to experimental condition (human tutor, virtual tutor, no tutoring) and whether the assessment was pre or post. Interrater reliabilities for two raters were high (counting only the open-ended items), with intraclass correlation coefficients ranging from .89 to .98, with averages for pre and post of .93 and .94, respectively. Internal reliabilities (Cronbach's alpha) were lower, ranging from $\alpha = .60$ to $\alpha = .89$ for both pre and post versions of the assessments, with averages for pre = .74 and post = .79. Scores used for outcome analysis were the averages across both raters.

Results

Several comparisons were made to test the hypotheses. To make comparisons, both standardized pre/post scores and *residual gain scores* compared groups on the average differences between their observed and expected scores. Gain differed markedly depending on where students started on the pretest, regardless of which group they belonged to. Students who started lower on the pretest gained more than students starting higher. This is often a sign of regression toward the mean where greater gain occurs for students starting lower regardless of actual learning. Regression toward the mean complicated the group comparisons for this study because the control students on average scored much lower on the pretest than students receiving tutoring. We believe the lower pretest scores for the control were primarily due to two factors:

1. Consented students (those whose parents returned signed permission forms) had higher pretest scores than nonconsented students. Pretest scores for nonconsented students were similar to the control group.

2. Schools choosing to participate as treatment groups in the study were not representative of the overall free and reduced lunch (FRL) percentage of the district. Boulder Language Technologies worked with BVSD officials to identify a set of schools to recruit. All classroom teachers for the targeted grades in those schools were recruited, and all of the teachers who agreed to participate were enrolled. In this particular study, those teachers who agreed to participate represented schools that had smaller percentages of FRL students. Schools with higher percentages of FRL students tend to have lower test scores, and more of these schools were in the control group.

When group comparisons were made, control students tended to gain more pre to post than tutored students simply because they started lower on the pretest. Residual gain scores and analysis of covariance (ANCOVA) were used for analysis to adjust for these differences in prescore (Rudestam & Newton, 1999). The residual gain score is the observed score minus the expected score in the scatter between pre and post; the expected score is the regression line for the scatter. It is used to compare

groups and has a mean of zero, with a scale representing standard deviation units.

Comparison Between Tutored Groups

The first hypothesis examined whether MyST and human-tutored groups were roughly equal to each other in pre/post gain. Students were randomly assigned within classrooms to tutoring conditions. Standardized gain for the human-tutored group ($M = 1.95$, $SD = 0.85$) was not significantly different than for the MyST-tutored group ($M = 1.75$, $SD = 1.03$), $t(150) = -1.31$, $p = .190$, $d = .18$. Residual gain for the human-tutored group ($M = 0.51$, $SD = 0.66$) was also not significantly different than for the MyST-tutored group ($M = 0.38$, $SD = 0.76$), $t(150) = -1.15$, $p = .250$, $d = .15$. Power analysis showed that for an effect size of $d = .15$, sample sizes of 600 students per group would be needed to reach significance at the .05 level with 80% power. The small effect size and lack of statistical significance support the first hypothesis that benefits of tutoring are roughly equal for human tutors and Marni in pre/post gain.

Comparison With Control Group

As stated, comparisons with the students in control classrooms were complicated by differences in pre-test scores. To adjust for these differences, comparisons were made with residual gain scores and an ANCOVA to test the second hypothesis that students in tutored groups gained more than students in the control group. Standardized gain scores showed a moderate difference between MyST ($M = 1.75$, $SD = 1.03$) and control ($M = 1.57$, $SD = 1.01$; $d = .18$) and a larger difference between the human ($M = 1.95$, $SD = 0.86$) and control ($d = .40$). Effect sizes for residual gain scores were calculated by the difference in means between groups divided by the pooled standard deviation for the residual gain distribution. A moderate effect size was observed for the comparison of MyST tutoring ($M = .38$, $SD = .76$) and control ($M = -.06$, $SD = .84$; $d = 0.53$) and a larger effect size for human tutoring ($M = .51$, $SD = .66$) and control ($d = 0.68$). A one-way analysis of variance (ANOVA) tested whether group means differed significantly on residual gain score. The main effect for tutoring was significant, $F(2, 1164) = 26.06$, $p < .001$. Post hoc tests showed significant differences between both tutoring groups and the control group, and no significant differences between the two tutoring groups.

An ANCOVA confirmed the findings from the analysis of residual gains. Like residual gain scores, ANCOVA also adjusts group means for differences in pretest. ANCOVA in this context gave almost identical results to the ANOVA using residual gains, $F(2, 1163) = 26.60$, $p < .001$. Comparisons of adjusted means were also nearly identical to effect sizes in residual gains for groups. ANOVA and ANCOVA tests support the second hypothesis that tutored groups gain significantly more from pre to post than students in the control group.

Gain was also assessed as a function of prescore. Group comparisons divided the prescore distribution for the tutored group into five equal parts. All groups showed higher gain for the lower prescore blocks.

The use of hierarchical models allows for partitioning of error between students and classrooms, and quantifying how much total

variability is due to each level. Estimates of classroom variability, calculated with all students in the classroom, equaled 46%. Hypothesis testing for classroom effects showed significant effects for both MyST compared with control, $t(60) = 2.5$, $p = .014$, and human compared with control, $t(60) = 3.0$, $p = .004$. These results from hierarchical models also support the second hypothesis that tutored groups gain more from pre to post than the control group.

Component Evaluation

In order to evaluate the performance of the speech-processing components, student utterances for a subset of the assessment data were manually transcribed and parsed into frames to give the reference data to compare against. ASR performance is typically expressed as a word error rate (WER), which is the sum of word deletion, insertion, and substitution errors divided by the number of words in the reference string (from human transcriptions). The speech recognizer vocabulary size was 6,235 words. The WER for the assessment sessions was 41.4%.¹ This is a large WER, and would not be viable for many applications. The system performed well even with the high WER because the accuracy of extraction of frame elements (the key concepts being discussed) from student's speech remained relatively high, with an overall Recall = 79% and Precision = 82%. So 79% of the relevant information in the reference parses was correctly extracted from the ASR output. Of the information extracted, 82% of the elements were correct. These results indicate that many of the recognition errors were in information that was not relevant or redundant. Given the nature of QtA dialogs and the way spoken responses are used by the system, this level of extraction accuracy was sufficient to produce both engaging and effective dialogs, as indicated by students' responses to questionnaires and the learning gains.

Survey Results

A written survey was given to the students who participated in the 2010–2011 assessment. Measures were taken to avoid bias wherein students give overly positive answers to questionnaires including the following: (a) Written (vs. oral) surveys for students were administered, (b) students were verbally assured of anonymity, (c) questionnaires were anonymous in that students did not write their names on the survey, and (d) adults from the program did not directly observe or interfere with students while they completed the survey. The survey included questions that asked for ratings of student experience and impressions of the program and its usability. Three-point rating scales for survey items were keyed to each question. A typical question, such as *How much did Marni help with science?* had responses such as: *Did not help*, *helped some*, *helped a lot*. Items were written to reflect the reading level of the students. In general, students had positive experiences and impressions about the program. Across schools, 47% of students said they would like to talk with Marni after every science investigation, 62% said they enjoyed working with Marni “a lot,” and

¹ The performance of the ASR system was enhanced significantly over the course of the project, and WER on the assessment data is now 21%. However, the system and models were fixed at the start of the assessment to avoid confounding the evaluation results with improvements in the performance of the speech recognition system.

53% selected "I am more excited about science" after using the program. Only 4% felt that the tutoring did not help.

Teachers were asked for feedback to help assess the feasibility of an intervention using the system and their perceptions of the impact of the system. A teacher survey was given to all participating teachers directly after their students completed tutoring. Teachers were assured anonymity in their responses both verbally and in written form. The questionnaire contained 22 rating items as well as nine open-ended questions. The survey asked teachers about the perceived impact of using Marni for student learning and engagement, impacts on instruction and scheduling, willingness to potentially adopt Marni as part of classroom instruction, and overall favorability toward participating in the research project. Additionally, teachers answered items related to potential barriers in implementing new technology in the classroom. Of the responding teachers, 100% said that they felt it had a positive impact on their students, they would be interested in the program if it were available, and they would recommend it to other teachers. In addition, 93% said that they would like to participate in the project again. Furthermore, 74% indicated that they would like to have all of their students use the system (not just struggling students). They commented that students who used the system were more enthused about and engaged in classroom activities and that their participation in science investigations and classroom discussions benefitted students who did not use the system.

Conclusion

In the present article, we presented the motivation, design, and evaluation results for a conversational multimedia virtual tutor for elementary school science. The operating principles for the tutor are grounded in research from education and cognitive science. Speech, language, and character animation technologies play a central role because the focus of the system is on engagement and spoken explanations by students during spoken dialogs with a virtual tutor.

An assessment was conducted in schools to compare learning gains from human tutoring and MyST with business-as-usual classrooms. Both tutoring conditions had significantly higher learning gains than the control group. Although the effect size for human tutors versus control ($d = 0.68$) was larger than for MyST versus control ($d = 0.53$), statistical tests supported the hypothesis of no significant difference between the two.

After the assessment, surveys were collected from students and teachers that bear on the engagement and feasibility of the tutoring system. Following a series of tutoring sessions with Marni, the great majority of students reported that they enjoyed spending time working with her, that they felt that Marni helped them learn science, and that they felt more interested in science and more motivated to learn science than they had before using the system. Teachers reported that they would like to use MyST in the future to tutor all of their students and that they would recommend the program to other teachers.

One conclusion that we draw from this study is that current spoken dialog and character animation technologies can be combined with media to provide engaging and effective experiences for third-, fourth-, and fifth-grade students learning science. Students who used MyST interacted with Marni for 4–5 hr over the course of the 16 dialog sessions over an 8- to 10-week period. No

students dropped out of the study, and the large majority of students reported positive experiences. We believe that the QtA approach helped assure the student that Marni is listening to and understands what they are saying; this experience is fostered by dialog moves such as revoicing and marking that Marni produces. Dialogs based on QtA enable the tutorial dialog to proceed in a graceful way even when the system does not accurately interpret what the student said, because the system typically proceeds with a reasonable follow-up question, which the student accepts as a natural extension of the dialog.

The system described presents baseline results for one specific system based on a number of design decisions. Further work is needed to understand the effects of the individual features of the system. For example, we do not know the relative contribution of media in helping students visualize science and construct explanations, or the contribution of the dialog moves and questions that Marni generated, to the learning gains that occurred. We believe the MyST system provides a framework and infrastructure for conducting research on these questions. Planned future work will allow us to expand the context of the interaction from one-on-one tutoring to systems that support conversations in which a virtual tutor is able to mediate conversations among small groups of students. The virtual tutor will then be able to ask questions that help students build on each other's ideas to coconstruct explanations consistent with accurate mental models of the science.

References

- Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94*, 416–427. doi:10.1037/0022-0663.94.2.416
- Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education, 15*, 95–115.
- Beck, I., McKeown, M., Sandora, C., Kucan, L., & Worthy, J. (1996). Questioning the author: A yearlong classroom implementation to engage students with text. *The Elementary School Journal, 96*, 385–414. doi:10.1086/461835
- Bloom, B. (1956). *Taxonomy of educational objectives, Handbook I: The cognitive domain*. New York, NY: David McKay.
- Bloom, B. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4–16.
- Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98*, 182–197. doi:10.1037/0022-0663.98.1.182
- Cazden, C. B. (1988). *Classroom discourse: The language of teaching and learning*. Portsmouth, NH: Heinemann.
- Chi, M. (2009). Active–constructive–interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science, 1*, 73–105.
- Chi, M., Bassok, M., Lewis, M., Reimann, P., Glaser, R., & Alexander. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182. doi:10.1207/s15516709cog1302_1
- Chi, M., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.
- Chi, M., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471–533. doi:10.1207/s15516709cog2504_1

- Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19, 237–248.
- Craig, S., Gholson, B., Ventura, M., & Graesser, A. (2000). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education*, 11, 242–253.
- Driscoll, D., Craig, S., Gholson, B., Ventura, M., Hu, X., & Graesser, A. (2003). Vicarious learning: Effects of overhearing dialog and monologue-like discourse in a virtual tutoring session. *Journal of Educational Computing Research*, 29, 431–450. doi:10.2190/Q8CM-FH7L-6HJU-DT9W
- Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington DC: National Academy Press.
- FOSS. (n.d.). *About FOSS*. Retrieved from <http://www.fossweb.com>
- Gamoran, A., & Nystrand, M. (1991). Background and instructional effects on achievement on eighth-grade English and social studies. *Journal of Research on Adolescence*, 1, 277–300.
- Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22, 39–51.
- Hake, R. (1998). Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66, 64–74.
- Hausmann, R. G. M., & VanLehn, K. (2007a). Explaining self-explaining: A contrast between content and generation. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education* (pp. 417–424). Amsterdam, the Netherlands: IOS Press.
- Hausmann, R. G. M., & VanLehn, K. (2007b). *Self-explaining in the classroom: Learning curve evidence*. Paper presented at the 29th Annual Conference of the Cognitive Science Society, Mahwah, NJ.
- Horz, H., & Schnotz, W. (2010). Multimedia: How to combine language and visuals. *Language at Work—Bridging Theory and Practice*. Retrieved from <http://ojs.statsbiblioteket.dk/index.php/law/article/view/6200>
- King, A. (1989). Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology*, 14, 366–381. doi:10.1016/0361-476X(89)90022-2
- King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, 31, 338–368.
- King, A., Staffieri, A., & Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology*, 90, 134–152. doi:10.1037/0022-0663.90.1.134
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163–182. doi:10.1037/0033-295X.95.2.163
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.
- Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). *The persona effect: Affective impact of animated pedagogical agents*. Paper presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems, Atlanta, GA.
- Madden, N. A., & Slavin, R. E. (1989). Effective pullout programs for students at risk. In R. E. Slavin, N. L. Karweit, & N. A. Madden (Eds.), *Effective programs for students at risk* (pp. 52–72). Boston, MA: Allyn & Bacon.
- Mayer, R. (2001). *Multimedia learning*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139164603
- Mayer, R. (2005). Introduction to multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 1–16). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511816819.002
- McKeown, M., & Beck, I. (1999). Getting the discussion started. *Educational Leadership*, 57, 25–28.
- McKeown, M., Beck, I., Hamilton, R., & Kucan, L. (1999). *"Questioning the Author" accessibles: Easy access resources for classroom challenges*. Bothell, WA: The Wright Group.
- Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, 19, 177–213. doi:10.1207/S1532690XCI1902_02
- Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart machines in education: The coming revolution in educational technology* (pp. 169–234). Cambridge, MA: MIT Press.
- Murphy, P. K., & Edwards, M. N. (2005). *What the studies tell us: A meta-analysis of discussion approaches*. Paper presented at the American Educational Research Association, Montreal, Canada.
- Murphy, P., Wilkinson, I., Soter, A., Hennessey, M., & Alexander, J. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology*, 101, 740–764. doi:10.1037/a0015576
- Nass, C., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. Cambridge, MA: MIT Press.
- National Assessment of Educational Progress. (2005). *National and state reports in science: The nation's report card*. Jessup, MD: ED Pubs.
- Nystrand, M. (1997). *Opening dialogue: Understanding the dynamics of language and learning in the English classroom*. New York, NY: Teachers College Press.
- Osborne, J. (2010, April 23). Arguing to learn in science: The role of collaborative, critical discourse. *Science*, 328, 463–466.
- Palincsar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review of Psychology*, 49, 345–375. doi:10.1146/annurev.psych.49.1.345
- Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1, 117–175. doi:10.1207/s1532690xci0102_1
- Pine, K., & Messer, D. (2000). The effect of explaining another's actions on children's implicit theories of balance. *Cognition and Instruction*, 18, 35–51. doi:10.1207/S1532690XCI1801_02
- Reeves, B., & Nass, C. (1996). *The media equation*. New York, NY: Cambridge University Press.
- Roy, M., & Chi, M. (in press). The self-explanation principle. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning*.
- Rudestam, E., & Newton, R. (1999). *Your statistical consultant: Answers to your data analysis questions*. Washington DC: Sage.
- Soter, A. O., & Rudge, L. (2005). *What the discourse tells us: Talk and indicators of high-level comprehension*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.
- Soter, A., Wilkinson, I., Murphy, P., Rudge, L., Reninger, K., & Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research*, 47, 372–391. doi:10.1016/j.ijer.2009.01.001
- Topping, K., & Whiteley, M. (1990). Participant evaluation of parent-tutored and peer-tutored projects in reading. *Educational Research*, 32, 14–32. doi:10.1080/0013188900320102
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46, 197–221. doi:10.1080/00461520.2011.611369
- VanLehn, K., & Graesser, A. C. (2001). *Why2 Report: Evaluation of Why/Atlas, Why/AutoTutor, and accomplished human tutors on learning gains for qualitative physics problems and explanations*. Unpublished report, University of Pittsburgh CIRCLE group and the University of Memphis Tutoring Research Group.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, 31, 3–62. doi:10.1080/03640210709336984

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J., Shelby, R., Taylor, L., . . . Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, 15, 147–204.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wise, B., Cole, R., Van Vuuren, S., Schwartz, S., Snyder, L., Ngampatpong, N., . . . Pellom, B. (2005). *Learning to read with a virtual tutor: Foundations to literacy: Interactive literacy education: Facilitation literacy environments through technology*. Mahwah, NJ: Erlbaum.

Received December 15, 2011

Revision received November 19, 2012

Accepted December 10, 2012 ■

Call for Nominations

The Publications and Communications (P&C) Board of the American Psychological Association has opened nominations for the editorships of **History of Psychology**; **Journal of Family Psychology**; **Journal of Personality and Social Psychology: Personality Processes and Individual Differences**; **Psychological Assessment**; **Psychological Review**; **International Journal of Stress Management**; and **Personality Disorders: Theory, Research, and Treatment** for the years 2016–2021. Wade Pickren, PhD, Nadine Kaslow, PhD, Laura King, PhD, Cecil Reynolds, PhD, John Anderson, PhD, Sharon Glazer, PhD, and Carl Lejuez, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2015 to prepare for issues published in 2016. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations are also encouraged.

Search chairs have been appointed as follows:

- **History of Psychology**, David Dunning, PhD
- **Journal of Family Psychology**, Patricia Bauer, PhD, and Suzanne Corkin, PhD
- **JPSP: Personality Processes and Individual Differences**, Jennifer Crocker, PhD
- **Psychological Assessment**, Norman Abeles, PhD
- **Psychological Review**, Neal Schmitt, PhD
- **International Journal of Stress Management**, Neal Schmitt, PhD
- **Personality Disorders: Theory, Research, and Treatment**, Kate Hays, PhD, and Jennifer Crocker, PhD

Candidates should be nominated by accessing APA's EditorQuest site on the Web. Using your Web browser, go to <http://editorquest.apa.org>. On the Home menu on the left, find "Guests." Next, click on the link "Submit a Nomination," enter your nominee's information, and click "Submit."

Prepared statements of one page or less in support of a nominee can also be submitted by e-mail to Sarah Wiederkehr, P&C Board Search Liaison, at swiederkehr@apa.org.

Deadline for accepting nominations is January 11, 2014, when reviews will begin.

A Tutoring System That Simulates the Highly Interactive Nature of Human Tutoring

Sandra Katz and Patricia L. Albacete
University of Pittsburgh

For some time, it has been clear that students who are tutored generally learn more than students who experience classroom instruction (e.g., Bloom, 1984). Much research has been devoted to identifying features of tutorial dialogue that can explain its effectiveness, so that these features can be simulated in natural-language tutoring systems. One hypothesis is that the highly interactive nature of tutoring itself promotes learning—that is, the *interaction hypothesis*. Although reasonable and agreeing with much research, the interaction hypothesis raises the question of what linguistic mechanisms are involved: that is, which features of “highly interactive” dialogues trigger what processes that are conducive to learning? Our overall strategy in the research described in this article was to inform this question by identifying co-constructed discourse relations in tutorial dialogues whose frequency of occurrence predicts learning, identify the context in which these relations occur, and use this knowledge to formulate decision rules to guide automated dialogues. We used Rhetorical Structure Theory to identify and tag co-constructed discourse relations in a large corpus of physics tutoring dialogues. Our analyses suggest that the effectiveness of human tutoring might well lie in the language of tutoring itself. Moreover, the types of co-constructed discourse relations that predict learning seem to vary based on students’ ability level. We describe Rimac, a natural-language tutoring system that implements an initial set of decision rules based on these analyses. These rules guide reflective dialogues about the concepts associated with physics problems. Rimac is being pilot tested in high school physics classes.

Keywords: instructional dialogue, natural-language tutoring systems, Rhetorical Structure Theory

Educators and policy makers in the United States have looked to educational technology as a tool to increase students’ proficiency in math, science, reading, and other subject matter domains. For example, early in his administration, President Obama (2009) challenged developers of intelligent tutoring systems (ITSs) to develop “learning software as effective as a personal tutor” (para. 19). Apparently, Obama cast this challenge a bit too late. A recent meta-analysis of research comparing the effectiveness of human tutors with state-of-the-art ITSs showed that ITSs have already nearly caught up with human tutors (VanLehn, 2011), with effect sizes (d) of 0.76 for human tutoring and 0.79 for ITSs relative to

no tutoring (e.g., problem solving and reading, without feedback).¹ This comparison raises the bar for developers of ITSs. The challenge now is to develop automated tutors that can perform even better than human tutors with learners of all types.

Several researchers have proposed that the large effect sizes of human tutoring can be attributed to its highly interactive nature—that is, the high degree to which the student and tutor respond to and build upon each other’s dialogue moves (e.g., M. T. H. Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001;² Graesser, Person, & Magliano, 1995; van de Sande & Greeno, 2010). However, an important line of research conducted in the past few years to test this so-called *interaction hypothesis* showed that it is neither how much interaction takes place during tutoring that is important, nor the granularity of interaction—for example, whether the student and tutor discuss a step toward solving a problem or the substeps that lead to that step. Instead, what matters most is how *well* the interaction is carried out—for example, what content is addressed and how it is addressed in a particular dialogue context (e.g., M. Chi, VanLehn, Litman, & Jordan, 2010, 2011a, 2011b; Murray & VanLehn, 2006).

This important finding suggests that the key to building tutoring systems that surpass the effectiveness of human tutors is to specify

This article was published Online First September 9, 2013.

Sandra Katz and Patricia L. Albacete, Learning Research and Development Center, University of Pittsburgh.

The authors thank the Rimac project team—Stefani Allegretti, Michael Ford, Pamela Jordan, Kevin Krost, Michael Lipschultz, Diane Litman, Tyler McConnell, Scott Silliman, Elizabeth Spiegel, Christine Wilson, and Peter Wu—for their contributions. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A10063 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Sandra Katz, Learning, Research and Development Center, 3939 O’Hara Street, Pittsburgh, PA 15260. E-mail: katz@pitt.edu

¹ VanLehn’s (2011) review showed that the two sigma effect for human tutoring reported by Bloom (1984) testifies to the importance of a mastery learning standard and is not typical of human tutoring in general.

² Two important players in the field of tutoring research have the same last name and first initial. In our citations, we use M. T. H. Chi to refer to Michelene (“Micki”) T. H. Chi and M. Chi to refer to Min Chi.

what we mean by *effective interaction* and to formulate “policies for selecting the tutorial action at each microstep when there are multiple action options available” (M. Chi et al., 2011a, p. 87). Such “policies” have alternatively been called *pedagogical tutoring tactics* or *pedagogical decision rules*. We use the latter term here (*decision rules*, for short). As several developers of natural-language (NL) tutoring systems have argued, since tutorial dialogue is a form of discourse, defining effective interaction entails identifying the particular linguistic mechanisms that support learning during tutorial interaction (e.g., Boyer et al., 2010; Di Eugenio & Green, 2010; Pilkington, 2001; Ravenscroft & Pilkington, 2000). Decision rules can then be specified to guide the tutor in determining when and how to carry out these linguistic mechanisms.

This article describes the development of Rimac, a natural-language tutoring system that scaffolds students in acquiring a deeper understanding of the physics concepts and principles associated with quantitative physics problems. Rimac was designed to supplement instruction in physics tutoring systems such as Andes (e.g., VanLehn et al., 2005).³ Rimac is primarily engineered to implement decision rules that guide the automated tutor in carrying out two linguistic mechanisms that have been found to predict learning from human tutoring: tutors’ abstraction and specification of students’ dialogue contributions (e.g., Katz, Allbritton, & Connelly, 2003; Ward, Connelly, Katz, Litman, & Wilson, 2009). This finding is supported by a significant body of prior research that demonstrates that the formation of abstract schema (i.e., mental representations of learned material) promotes transfer (e.g., Gick & Holyoak, 1983, 1987; Leher & Littlefield, 1993; Reed, 1993; Salomon & Perkins, 1989). During tutoring, abstraction takes place when the tutor or student relates what his or her dialogue partner said to explain a more general concept or principle. For example, during physics tutoring, abstraction involves mapping the physical state presented in a problem to concepts and principles that explain that state or to a general script for solving that type of problem. Specification is the reverse and typically occurs when the tutor (or student) distinguishes between related concepts, instantiates a formula that represents a physics principle, applies a problem-solving script to the problem at hand, and so forth.

From a linguistic perspective, abstraction and specification are often implemented through hypernym/hyponym pairs of terms (Halliday & Hasan, 1976). For example, in the following exchange from a live tutoring session, the tutor specifies “velocity” (hypernym) in the student’s turn to “horizontal components of the velocity” (hyponym).

Example 1

Student: Velocity is in the same direction as acceleration so the ball is faster coming down.

Tutor: It [the ball] slows down going up, and it speeds up coming down—but all the time the *horizontal components of the velocity* stay unchanged. [italics ours]

However, sometimes abstraction and specification are implemented through semantic relations between speaker turns, with few or no lexical cues such as those shown in Example 1, and inference is required to detect these semantic relations. For example, in the following exchange, the student needs to infer that the

tutor’s phrase “change in velocity” abstracts over the student’s phrase “final velocity is larger than the starting velocity.”

Example 2

Tutor: How do we know that we have an acceleration in this problem?

Student: Because the final velocity is larger than the starting velocity, 0.

Tutor: Right—a *change in velocity* implies acceleration. [italics ours]

In addition to implementing decision rules to guide the automated tutor in abstracting and specifying students’ dialogue turns, Rimac also simulates a few other linguistic processes that commonly occur during physics tutoring—most notably, joint construction of conditional reasoning relations, as we will illustrate presently (Louwerse, Crossley, & Jeuniaux, 2008).

Several studies have shown that data-driven machine learning techniques such as reinforcement learning can be applied to logged interactions from natural-language tutoring systems in order to derive decision rules to guide tutorial interaction (e.g., Beck, Woolf, & Beal, 2000; M. Chi et al., 2010, 2011a, 2011b; Murray & VanLehn, 2006). Evaluations of ITSs that implement these rules have found that these systems significantly outperform counterpart systems that carry out random policies—for example, “eliciting” a problem-solving step or dialogue goal from the student sometimes, “telling” the student that step or goal at other times, without clear guidelines about what to do when. Some rule-driven tutoring systems have also outperformed systems that implement “fixed” tutoring policies (e.g., Murray & VanLehn, 2006)—for example, responding to students’ help requests with increasingly directive feedback such as prompt first, then hint, then teach relevant background knowledge, and then (if all else fails) tell the student what to do (the so-called *bottom out hint*).

Although this research demonstrates the promise of automated methods for deriving effective decision rules to guide tutorial dialogue, it also shows that the process is both difficult and costly. As M. Chi et al. stated, “Finding effective tutorial tactics is not easy” (M. Chi, Jordan, VanLehn, & Litman, 2009, p. 197). In addition, the decision rules that stem from this approach are highly domain specific and difficult to interpret. Take, for example, one decision rule that M. Chi et al.’s (2011a) reinforcement-learning-based system defined for “elicit versus tell”—that is, should a tutor prompt the student for domain content at a particular point in a dialogue or tell the student that content?

Rule 6 suggests that when the next dialogue content step is difficult (StepSimplicityPS is 0), the ratio of physics concepts to words in the tutor’s turns so far is high (TuConceptsToWordsPS is 1), and the tutor has not been very wordy during the current session (TuAvgWordsSesPS is 0), then the tutor should tell. (p. 96)

On the one hand, finely nuanced rules such this one have the benefit that researchers using conventional experimental methods to test hypothesized decision rules could not predict these rules in

³ Rimac is the name of a river whose source is in the Andes. Its name is a Quechua word meaning *talking*; hence, the nickname for Rimac, *talking river*. We thus considered the name Rimac to be well suited for a dialogue system that could be embedded within the Andes tutoring system.

the first place. Similar observations have been made of the use of automated approaches to identify linguistic features of tutorial dialogue that predict learning, such as hidden Markov models (e.g., Boyer et al., 2010). On the other hand, rules such as this are cryptic and complex to implement, as the researchers have acknowledged.

In developing natural-language dialogues for Rimac, we strove to specify decision rules that were supported by preliminary empirical research, more intuitive than those illustrated previously, and readily implementable using a common framework for generating NL dialogues, which we will describe presently. Consequently, we took a more conventional approach. We first performed correlational analyses to identify specific relations between tutors' and students' dialogue moves in a large corpus of human-tutored physics dialogues that predict student learning gains from pretest to posttest. We then examined the context in which these relations typically occur and formulated decision rules that specify these contextual conditions. We implemented these rules within Rimac and are currently evaluating the system to determine if it outperforms a less interactive, less rule-driven tutoring system control.

In the next section, we situate Rimac in a framework of tutoring research that highlights the need for effective decision rules to guide natural-language dialogue systems. In keeping with the theme of this special issue of the *Journal of Educational Psychology*, we then describe the empirical research that we conducted to derive decision rules to guide abstraction, specification, and other commonly occurring relations between students' and tutors' dialogue turns, particularly during physics tutoring, and illustrate how we implemented these rules within Rimac.

Cooperative Execution During Scaffolding

The most intensive interaction during human one-on-one tutoring takes place during *scaffolding*, which M. T. H. Chi et al. (2001) defined as follows:

[A] scaffolding move is a kind of *guided prompting* that pushes the student a little further along the same line of thinking, rather than telling the student some new information, giving direct feedback on a student's response, or raising a new question or a new issue that is unrelated to the student's reasoning The important point to note is that scaffolding involves *cooperative execution* or *coordination* by the tutor and the student (or the adult and child) in a way that allows the student to take an increasingly larger burden in performing the skill. (p. 490).

The nexus of scaffolding lies in the fourth step of Graesser et al.'s (1995) "five-step dialogue frame" (p. 504) to describe the cyclic nature of tutorial interaction:

- Step 1. Tutor asks question.
- Step 2. Student answers question.
- Step 3. Tutor gives short feedback on the quality of the answer.
- Step 4. Tutor and student collaboratively improve the quality of the answer.
- Step 5. Tutor assesses student's understanding of the answer.

As Graesser et al. (1995) and others (e.g., VanLehn et al., 2007) have noted, understanding Step 4 of this frame—that is, scaffolding to improve the student's answer—could hold the key to understanding why human tutoring is so effective.

M. T. H. Chi et al.'s (2001) definition of scaffolding names two linguistic mechanisms that drive it: *coordination* and *cooperative execution*. We consider coordination first, because more research has been devoted to describing it. *Coordination* refers to the ways in which the tutor and student "stay on the same page"—that is, "grounding" the conversation, by acknowledging their dialogue partner's moves, negotiating the meaning of terms, and sharing knowledge (Clark & Schaefer, 1989; VanLehn, 2011). Coordination can also be supported by various forms of verbal alignment, such as lexical cohesion (e.g., word repetition, synonymy, paraphrase), and syntactic (word order) alignment (Garrod & Pickering, 2004). When the student hears his words (or word order) echoed in the tutor's turn, the student knows that the tutor understood what he or she said. Several studies have shown that the degree of lexical and syntactic cohesion (alignment) during tutoring predicts learning (e.g., Litman & Forbes-Riley, 2006; Steinhäuser et al., 2011; Ward & Litman, 2008, 2011), in addition to potentially enhancing coordination.

Cooperative execution refers to the joint construction of a line of reasoning. According to VanLehn (2011), cooperative execution takes place as tutors prompt students to continue a line of reasoning, indicate who should continue the execution, and accept the student's reasoning (p. 211). Our observations of tutorial dialogues reveal that cooperative execution during scaffolding involves more than these dialogue management processes; it also involves co-construction of the parts of an emerging line of reasoning or explanation. The analyses described in the Method section were motivated by our hypothesis that tutoring researchers need to formally describe these co-constructed dialogue moves and determine which types of moves support learning in order to develop natural-language dialogue systems that are as effective, or even more effective, than human tutors.

A Linguistic Framework to Describe Cooperative Execution

Rhetorical Structure Theory (RST) is a theoretical linguistic framework that specifies types of logical and functional relationships between parts of text and spoken discourse, including various types of abstraction and specification relations. Mann and Thompson (1988), who developed RST, argued that "it describes the relations among text parts in functional terms, identifying both the transition point of a relation and the extent of the items related" (p. 271). Functional and logical relationships between parts of spoken and written discourse go by many names, including *rhetorical relations*, *coherence relations*, and *discourse relations* (Hovy, 1990). We use the latter term here.

Table 1 defines and illustrates the set of abstraction/specification relations, and other discourse relations, which we manually tagged in a corpus of human tutorial dialogues in order to determine which co-constructed relations predict learning and are thereby most important to simulate in Rimac. For example, a student applies the equation for acceleration; the tutor then says something general about acceleration (e.g., "Acceleration is a vector and hence has direction as well as magnitude."). In RST, this is a jointly constructed *instance:abstract* discourse relation. To take another example, the tutor describes a set of conditions that apply to a given physical situation—for example,

Table 1
Discourse Relations Tagged in the Dialogue Corpus

Relation and definition (S = speaker)	Example
Abstraction/specification relations	
Abstract:instance (instance:abstract): S2 instantiates the abstraction stated by S1, or S2 abstracts over the information presented by S1.	<p><i>Tutor:</i> How can the acceleration be 0 if there are forces on it?</p> <p><i>Student:</i> The sum of the forces equal 0 for there to be no acceleration.</p> <p><i>Tutor:</i> That's exactly right. The weight and the normal force are (in this case) equal and opposite.</p> <p><i>Explanation:</i> "In this case" (as the tutor says), the weight and normal force being equal and opposite represent an instance of the abstraction "sum of forces equal 0."</p>
Set:member (member:set): S2 presents a member of the set referred to by S1, or S2 names the set to which an item mentioned by S1 belongs.	<p><i>Tutor:</i> What does the problem ask for?</p> <p><i>Student:</i> The magnitude of the acceleration</p> <p><i>Tutor:</i> What type of acceleration?</p> <p><i>Student:</i> Average</p> <p><i>Explanation:</i> The tutor refers to acceleration as a set and prompts for a member of that set; the student gives the type of acceleration asked for in the problem.</p>
Whole:part (part:whole): S2 names a part of an object that S1 referred to, or S1 names a part of an object named by S2. (In physics, "parts" are often vector components or the specific forces acting on an object.)	<p><i>Student:</i> Acceleration would be plus.</p> <p><i>Tutor:</i> Right, the x component of the acceleration would be plus.</p> <p><i>Explanation:</i> The student names a vector (acceleration); the tutor refers to a specific component of that vector.</p>
Process:step (step:process): S2 presents a step that follows from the process or line of reasoning described by S1, or S2 describes the line of reasoning that leads to the step described by S1.	<p><i>Student:</i> The acceleration is 0.</p> <p><i>Tutor:</i> So then $m \cdot a = 0 = F_{\text{net}} = T - W$ and hence $T = W$.</p> <p><i>Explanation:</i> The student gives a step in a line of reasoning; the tutor expands the line of reasoning (process) that follows from that step.</p>
Object:attribute (units, direction, magnitude): S1 names an object or value; S2 specifies a property of that object—in particular, its units, direction, or magnitude.	<p><i>Student:</i> Velocity is 14.</p> <p><i>Tutor:</i> Right, 14 m/s.</p> <p><i>Explanation:</i> The student provides a value for velocity; the tutor specifies its units.</p>
Term:definition (definition:term): S2 defines a term mentioned by S1, or S2 labels a statement by S1 with an appropriate term.	<p><i>Tutor:</i> What is the definition of the average acceleration (in words or in mathematics)?</p> <p><i>Student:</i> $A = (V_f - V_o)/T_f - T_o$.</p> <p><i>Explanation:</i> The tutor prompts the student to define average acceleration; the student does so.</p>
General:specific (specific:general): S2 names a state, object, or action that is related to the content in S1 but is more specific, or S2 is more general than the state, object, or action referred to in S1. Applies when none of the preceding relations apply.	<p><i>Student:</i> Average acceleration can vary.</p> <p><i>Tutor:</i> Right; it can go up above the average and down below it.</p> <p><i>Explanation:</i> The tutor specifies how acceleration can vary.</p>
Other commonly occurring relations in physics tutoring	
Condition:situation (situation:condition): (a) S1 presents a condition or set of circumstances, and S2 states the situation that stems from or coincides with those conditions, or (b) S1 presents a situation, and S2 states the conditions or circumstances that explain that situation.	<p><i>Tutor:</i> When do kinematics equations apply?</p> <p><i>Student:</i> When the acceleration is constant.</p> <p><i>Explanation:</i> This relation could be stated in conditional form: if acceleration is constant, then the kinematics equations apply.</p>
Compare: S2 compares an object, situation, or value referred to by S1 with some other object, situation, or value.	<p><i>Tutor:</i> What is the net force that the air bag imparts to the driver?</p> <p><i>Student:</i> Equal to the force the driver applies to the airbag.</p> <p><i>Tutor:</i> Same direction?</p> <p><i>Student:</i> No, opposite direction.</p> <p><i>Explanation:</i> The tutor prompts the student to compare the value and direction of two.</p>

"A car is moving to the right and is suddenly stopped"—and then prompts the student to state the situation that follows from this set of conditions—for example, that the car's acceleration is to the left. This is a co-constructed *condition:situation* (conditional) relation. Any relation can be delivered didactically, by the tutor or student, instead of interactively, as in these examples. For example, the tutor could have stated the same conditional relation didactically as follows: "Since the car is moving to the right and is suddenly stopped, its acceleration is to the left." However, we focused our investigation on the potential relationship between co-constructed discourse relations

and learning because these relations realize cooperative execution during scaffolding.

Method

To reiterate, our goals in the analyses described in this section were to (a) determine if the frequency of particular types of co-constructed discourse relations (those described and illustrated in Table 1) predict learning, and whether this varies by student ability level, and (b) formulate decision rules that specify the context in which those

discourse relations predicting learning occurs, so that these rules can guide student–tutor interaction in a NL tutoring system (Rimac). Toward these aims, we coded all instances of co-constructed discourse relations in a large corpus of human-tutored physics dialogues. The dialogue corpus and our approach to coding identified relations are described in this section.

Dialogue Corpus

A well-known problem in physics education is that many students learn to apply scripts for solving particular types of problems and succeed in college-level physics courses; however, they nonetheless leave these courses without understanding fundamental physics concepts and principles (Halloun & Hestenes, 1985). Reflective discussions following problem-solving exercises encourage students to think about the concepts and principles associated with quantitative problems, often by changing some aspect of the problem and prompting the student to consider how the answer would change, as illustrated in Table 2. Several studies have demonstrated the instructional benefits of reflection on problem-solving exercises (e.g., Collins & Brown, 1986; Katz, Connelly, & Wilson, 2007; Katz et al., 2003; Lee & Hutchison, 1998; Tchetagni, Nkambou, & Bourdeau, 2007; Ward & Litman, 2011).

The dialogue corpus that we analyzed stems from previous research in which we compared the effectiveness of human-guided reflective discussions about physics problems solved within the Andes physics tutoring system (VanLehn et al., 2005) with static text explanations

and a no-dialogue control. We summarize the data collection procedures that produced the dialogue corpus in this section. More details about the study can be found in Katz et al. (2003).

Students who were taking an introductory physics course at the University of Pittsburgh first took a physics pretest, with nine quantitative and 27 qualitative physics problems. Following the pretest, students reviewed a workbook chapter developed for the experiment and then received training on using Andes. There were three conditions: one in which students received reflection questions and interacted with a human tutor via a chat interface; a second reflection condition in which students were asked the same set of reflection questions but received a static text explanation as feedback after they responded to these questions; and a third, a control condition in which students were not asked reflection questions but solved more problems than students in the other two conditions to control for time on task. There were 15 students in the static text and control conditions and 16 students in the human-tutored condition. In the correlational analyses discussed here, we only analyzed data from the human-tutored condition, since we were interested in modeling effective aspects of human tutorial dialogue.

Students in each condition began by solving a problem in Andes. After completing the problem, students in both the static feedback and human-tutored conditions were presented with a conceptually oriented reflection question, as illustrated in Table 2. Reflection questions such as the one shown in Table 2 are not part of Andes; they were added for the experiment. After a student in the human-tutored condition entered a response to the reflection question, the student engaged in a typed dialogue with his or her tutor via a simple chat interface. This dialogue continued until the tutor was satisfied that the student understood the correct answer to the question.

Between three and eight reflection questions were asked per problem solved in Andes for a total of 12 problems. After completing these problems and their corresponding reflective dialogues, students took a posttest that was isomorphic to the pretest, and the test order was counterbalanced. The main finding of the study was that students who answered reflection questions learned more than students in the no-reflection control, who solved more Andes problems (Katz et al., 2003). Consistent with authors of several other studies who found a null effect for the interaction hypothesis, we did not observe a significant difference between the static feedback and human-tutored conditions (VanLehn, 2011; VanLehn et al., 2007). However, the human-tutored dialogue corpus revealed abundant instances of highly interactive, cooperative execution during scaffolding episodes—specifically, exchanges in which the tutor incorporated parts of the student's turn, built on the student's turn, and so on (e.g., Table 2)—or less frequently, the student did the same with respect to a preceding tutor turn. Hence, we deemed this corpus well-suited for exploring correlations between interactivity and student learning outcomes.

The dialogue corpus is sizeable. Among the 16 students in the human-tutored condition (four men, 12 women), 15 completed all 60 reflection question dialogues with a human tutor; one student participated in 53 dialogues, producing a total of 953 reflective dialogues. There were a total of 2,218 student turns and 2,135 tutor turns across dialogues. The average number of turns per reflective dialogue was 4.6, ranging from 2.1 turns for simple reflection questions to 11.4 turns for the most complex questions. All dia-

Table 2

Example of a Reflective Dialogue Between a Human Tutor and Student

Problem: In the figure below, each of the three strings exerts a tension force on the ring as marked. Use the labels S1, S2, and S3 to refer to the three strings. Find the components of the net force acting on the ring.



Reflection question: What if I now told you that this ring has an acceleration. If you knew the mass of the ring (3 kg), how would you solve for the acceleration?

Student: $73.2 - 3^*a$; $100 - Fw = 3^*a$. Is this right; how would the acceleration be the same for both?

Tutor: You have to keep the a_x and a_y distinguished. They are two completely independent numbers that (together with a_z) specify your acceleration vector. You don't try to boil them down to one number. It's as if I told you, "To get to my house, you go 3 blocks north and 5 blocks east," and you said, "Ah, so you just go 8 blocks"—the two numbers together are the vector; they don't "boil down" to one number. OK?

Student: But can't it only have one acceleration?

Tutor: It does have only one acceleration, but that acceleration is a vector and it takes 3 numbers to write it down. You need to review vectors in some detail; a_x , a_y , and a_z together specify the acceleration vector.

Note. This example problem is part of the Andes Physics Tutor system, which was developed at Arizona State University and the University of Pittsburgh with support from the Pittsburgh Science of Learning Center, National Science Foundation Award SBE-0836012, and Office of Naval Research Grant N00014-96-1-0260 and is available at <http://www.andestutor.org>

logue examples presented in this article stem from this tutoring corpus, unless otherwise noted.

Coding Scheme

Within each reflective dialogue, all student and tutor turns were first manually parsed into clauses. We then searched for co-constructed discourse relations at the exchange level—that is, between a tutor's dialogue turn and the subsequent student turn, or the reverse. We coded these relations at two levels of analysis: abstraction level type, and discourse relation type.

Abstraction Level Type

At the coarsest level, we tagged the level of abstraction of each exchange in which a discourse relation was co-constructed. Four codes distinguish these levels of abstraction, as described in the following. Code abbreviations are shown in parentheses.

Specific-to-general (spec:gen). This code refers to abstraction, which happens in two main ways. The first type is when the second speaker refers to a more general concept, principle, or value than one that the first speaker referenced in his dialogue turn. For example, in the following exchange, the tutor refers to speed, and the student classifies speed as a scalar quantity:

Example 3

Tutor: Since the question asked about SPEED, suppose we had found v_y to be negative. Should we include the minus sign when giving the speed?

Student: I would say no because speed is scalar and doesn't include direction.

In the second type of abstraction, the second speaker refers to a physics principle that explains or is illustrated by problem-specific content in the first speaker's turn. For example, in the following exchange, the tutor prompts the student to apply a principle about the relationship between acceleration and velocity to the bullet in the case at hand:

Example 4

Reflection question: The bullet is travelling to the right. What direction is its acceleration?

Student: To the left because it is making the bullet slow down.

Tutor: Good—when something is slowing down, its acceleration has a component opposite to its velocity.

General-to-specific (gen:spec). This code refers to specification, which is the inverse of abstraction and also happens in two main ways. The first type is when the second speaker refers to a more specific concept, principle, or value than the one to which the first speaker referred. For example, in the following exchange, the tutor asks for the forces on a climber, and the student names two types of forces:

Example 5

Tutor: What are the *forces* on her?

Student: Her *weight* and the *tension* of the rope. [italics ours]

The second main type of specification is when the second speaker instantiates a principle or concept to which the first speaker refers. For example, in the following exchange, the student carries out the tutor's directive to apply Newton's second law to the current problem:

Example 6

Tutor: Now use Newton's second law and find [the climber's] acceleration—a number and units; show me the symbols (the algebra).

Student: $39/55 = a$, $a = .71 \text{ m/s}^2$ downward.

Specific (spec). This code refers to cases in which the student and tutor are both speaking at the same level of abstraction, typically in reference to a particular problem. For example, in the following exchange, the tutor and the student refer to the bungee in the current problem. The tutor explains the situation that would result from the student's erroneous claim via a co-constructed conditional relation:

Example 7

Student: The only force acting on the bungee is the weight of the person.

Tutor: If that were true, the bungee would accelerate downward!

General (gen). This code refers to cases in which the student and tutor both speak at an abstract level, referring to principles, laws, definitions, and so forth that are not directly tied to a particular problem. For example, in the following exchange, the tutor and student step outside of the context of the current problem (about a falling hailstone) to discuss the difference between distance and displacement, in this comparison relation:

Example 8

Tutor: Is there a difference between displacement and distance?

Student: The displacement can have either value [+ or −], but distance is only +.

Table 3 presents the mean and standard deviation of abstraction level tags across subjects.

Discourse Relation Type

At a finer level of analysis, we tagged the dialogue corpus for the particular types of abstraction and specification relations defined and illustrated in Table 1, in addition to two other commonly occurring discourse relations in physics tutoring dialogues—conditional reasoning statements and comparisons. Most of these discourse relations are bidirectional (e.g., set:member, member:

Table 3
Mean Frequency of Abstraction Level Tags Across Tutored Subjects (N = 16)

Abstraction level	Mean	SD
Specific-to-general	14.13	4.83
General-to-specific	37.31	15.12
Specific	3.31	2.18
General	11.06	5.31

set); the exceptions are object:attribute and compare. We tagged bidirectional relations separately (e.g., we treated set:member and member:set as individual relations) and also treated each of the object:attribute categories as a separate relation. Hence, overall, there are 17 discourse relations in our coding scheme.

The basic unit of analysis at the discourse relation level is one of these codes, specified in two ways. First, we specify the *direction* of the co-constructed relation in the exchange—that is, does the tutor (T) start the relation and then the student (S) completes it, or the reverse? The former is indicated by T-S before the discourse relation name, and the latter by S-T—for example, S-T set:member represents a set:member relation that the student initiates and the tutor completes; and T-S abstract:instance represents an abstract:instance relation that the tutor initiates and the student completes. To illustrate, in the example shown in Table 1 for set:member, the second exchange (T: What type of acceleration? S: Average) would be tagged as T-S set:member.

The second way in which we modify discourse relation tags is by indicating whether the second turn in a tagged relation was prompted, via a question, or initiated by the second speaker. Prompted relations, such as the one for set:member, are unmodified—that is, T-S set:member means that the tutor prompted the student to provide a member of a named set, as in the preceding example about “type of acceleration.” Initiated relations are flagged as elaborations (elab), because the second speaker is adding information to what the first speaker said. To illustrate, in the example for abstract:instance shown in Table 1, the tutor elaborates on the student’s turn, by instantiating the student’s abstract statement:

Example 9

Student: The sum of the forces equals 0 for there to be no acceleration.

Tutor: That’s exactly right. The weight and the normal force are (in this case) equal and opposite.

This relation would be tagged as S-T elab(abstract:instance) to indicate that the tutor elaborated on the student’s statement via an abstract:instance relation. Instantiation is signaled by the tutor’s phrase “in this case.”

In addition to prompted and initiated variants of discourse relations, in both directions (S-T and T-S), we included three types of aggregate variables in our analyses. One aggregate variable includes the four prompted and initiated (elaborated) forms of a discourse relation. For example, the aggregate variable *whole:part* represents:

S-T whole:part + T-S whole:part + S-T elab(whole:part) + T-S elab(whole:part).

The second type of aggregate variable includes the four forms of the first relation, plus the four forms of its inverse. For example, the following formula represents *all-whole:part-bd*, where *bd* means bidirectional, for a particular relation (e.g., whole:part and part:whole, each consisting of the four forms shown in the formula):

[S-T whole:part + T-S whole:part + S-T elab(whole:part) + T-S elab(whole:part)] + [S-T part:whole + T-S part:whole + S-T elab(part:whole) + T-S elab(part:whole)].

The third type of aggregate variable includes the summation of all initiated elaborations. Specifically, T-S elab is the summation of student elaborations on the tutor’s previous turn, for all base

relations (e.g., whole:part, set:member); S-T elab is the summation of tutor elaborations of the student’s previous turn, for all base relations; and all-elab-bd = T-S elab + S-T elab.

Table 4 summarizes the means and standard deviation of discourse relation tags and aggregate tags across subjects.

Data Analysis

We conducted correlational analyses between the frequency of abstraction level codes, discourse relation codes, and three measures of student learning: overall gain score from pretest to posttest, gain score on qualitative test items, and gain score on quantitative test items. We conducted these analyses taking the 16 tutored students as a whole and separately for low and high pretest students, as classified according to a median split. There were seven high pretest students and nine low pretest students. These numbers are uneven because the two pretest scores in the middle of the distribution were identical; both students who had these scores were assigned to the low pretest group. We divided students into these ability groups in order to investigate whether better prepared students (high pretesters) might benefit from co-constructing dif-

Table 4
Mean Frequency of Discourse Relation Tags Across Tutored Subjects (N = 16)

Discourse relation variable or aggregate variable	Mean	SD
Abstract:instance	9.63	5.28
Instance:abstract	3.50	2.34
All-abstract:instance-bd	13.13	6.02
All-compare	3.19	1.83
Term:definition	3.00	2.25
Definition:term	0.13	0.34
All-term:definition-bd	3.13	2.22
Object:attribute-units	1.63	2.06
Object:attribute-direction	4.06	2.41
Object:attribute-sign	0.19	0.40
Object:attribute-magnitude	0.69	0.79
All-object-attribute	6.56	3.76
Process:step	0.56	0.63
Step:process	3.00	2.34
All-process:step-bd	3.56	2.31
Set:member	0.88	1.50
Member:set	2.00	1.67
All-member:set-bd	2.88	2.58
Whole:part	2.88	1.78
Part:whole	0.44	0.73
All-part:whole-bd	3.31	1.99
Circumstance:situation	13.00	6.79
Situation:circumstance	9.19	2.90
All-circumstance:situation-bd	22.19	6.93
Gen:spec	3.31	1.99
Spec:gen	0.75	1.07
All-gen:spec-bd	4.06	2.24
T-S elab	1.00	1.10
S-T elab	22.13	12.15
All-elab-bd	23.13	12.41

Note. Aggregate variables include modified forms of the base relation (e.g., whole:part) as described in the text. Gen = general; Spec = specific; T = tutor; S = student; bd = bidirectional; T-S elab = summation of student elaborations on the tutor’s previous turn, for all base relations; S-T elab = summation of tutor elaborations of the student’s previous turn, for all base relations; all-elab-bd = T-S elab + S-T elab.

ferent types of discourse relations with their tutor than less well-prepared students (low pretesters).

The results of these analyses are presented in the next section. We then describe the decision rules that stem from these findings.

Results and Discussion

Discourse Relations That Predict Learning: All Students Considered Together

Correlations for the subject pool taken as a whole ($N = 16$) are displayed in Table 5. To save space, we only discuss significant findings ($p \leq .05$) for all three types of gain.

Overall gain. The frequency of three discourse relations predicted overall gain: (a) various forms of the whole:part relation [S-T elab(whole:part) and two aggregate variables: whole:part and all-whole:part-bd], (b) S-T situation:condition relations, in which the student prompts the tutor to specify the conditions under which a physical situation occurs and the tutor replies accordingly, and

(c) various forms of the step:process relation [S-T elab(step:process) and the aggregate variable step:process], in which one dialogue partner provides the steps in a line of reasoning that stem from, or lead to, a step in his partner's turn, for example:

Example 10

Reflection question: How do we know that we have an acceleration in this problem?

Student: Because of gravity pulling down.

Tutor: The force due to gravity produces a net force and thus an acceleration.

In this exchange, the tutor provides the line of reasoning that follows from the student's response (gravity \rightarrow existence of a net force \rightarrow existence of acceleration), via an S-T elab(step:process) relation.

Qualitative gain. Generalizations predicted learning of a qualitative (conceptual) nature; a trend was also found for generalization and overall gain. This is not surprising, given that generalizations typically address physics concepts, laws, and principles. As with overall gain, various forms of the step:process relation also predicted qualitative gain across subjects [S-T elab(step:process) and two aggregate variables: step:process and all-process-step-bd]. In addition, a particular type of generalization predicted qualitative gain: S-T elab(member:set), in which the tutor elaborates on a student turn by stating the set to which an object that the student referred to belongs:

Example 11

Reflection question: How do we know that we have an acceleration in this problem?

Student: Because it is a free fall problem so gravity is at work.

Tutor: Gravity is a type of acceleration.

Quantitative gain. The "spec" abstraction level type, representing exchanges in which the tutor and student refer to the current problem, negatively correlated with quantitative gain. However, two particular forms of specification strongly predicted quantitative gain: S-T elab(set:member), in which the tutor states a member of a set that the student referred to, and S-T elab(whole:part), which typically reflects exchanges in which the tutor specifies the components of a vector that the student mentioned or the applied forces on an object that the student mentioned:

Example 12

Student: $(\text{String1} + \text{String2})/g = \text{mass of plane}.$

Tutor: It would be $(F_{1,y} + F_{2,y})/g = \text{mass, OK?}$

Discourse Relations That Predict Learning Among Low Pretest Students

Correlations for low pretest students ($N = 9$) are displayed in Table 6. We again focus our discussion on significant findings ($p \leq .05$) for all three types of gain.

Overall gain. Student generalizations over the tutor's turn positively correlated with low pretesters' overall gain score; how-

Table 5
Correlations for All Students Considered Together ($N = 16$)

Abstraction level and discourse relations	Mean	SD	R	p
Overall gain				
Abstraction level: [spec:gen]	14.13	4.829	.450	.081
Discourse relations				
S-T elab(step:process)	1.56	1.365	.646	.007**
step:process	3.00	2.338	.582	.018
S-T elab(member:set)	0.94	0.680	.667	.005**
S-T elab(whole:part)	1.00	1.366	.524	.037
whole:part	2.88	1.784	.528	.035
all-part:whole-bd	3.31	1.991	.553	.026
S-T situation:condition	0.44	0.814	.531	.034
[definition:term]	0.13	0.342	-.485	.057
[all-proc:step-bd]	3.56	2.308	.473	.064
Qualitative gain				
Abstraction level: spec:gen	14.13	4.829	.516	.041
Discourse relations				
S-T elab(step:process)	1.56	1.365	.653	.006**
step:process	3.00	2.338	.591	.016
all-proc:step-bd	3.56	2.308	.527	.036
S-T elab(member:set)	0.94	0.680	.558	.025
[T-S elab(term:definition)]	0.06	0.250	.469	.067
[definition:term]	0.13	0.342	-.443	.086
[S-T step:process]	0.06	0.250	.469	.067
[whole:part]	2.88	1.784	.463	.071
[all-part:whole-bd]	3.31	1.991	.457	.075
[S-T situation:condition]	0.44	0.814	.487	.056
Quantitative gain				
Abstraction level: spec	3.31	2.182	-.530	.035
Discourse relations				
S-T elab(set:member)	0.06	0.250	.740	.001**
S-T elab(whole:part)	1.00	1.366	.675	.004**
[T-S instance:abstract]	0.38	0.500	-.493	.052
[Object:attribute-magnitude]	0.69	0.793	-.467	.068
[Process:step]	0.56	0.629	-.445	.084
[S-T elab(member:set)]	0.94	0.680	.443	.086
[all-part:whole-bd]	3.31	1.991	.452	.079

Note. Trends are indicated by brackets. Gen = general; Spec = specific; T = tutor; S = student; bd = bidirectional; elab = elaborated; proc = process.

** $p < .01$.

Table 6
Correlations for Low-Pretest Students ($N = 9$)

Abstraction level and discourse relations	Mean	SD	R	p
Overall gain				
Abstraction level				
T-S spec:gen	4.44	3.005	.671	.048
S-T spec	1.67	1.732	-.719	.029
Discourse relations				
S-T situation:condition	0.67	1.000	.679	.044
[S-T elab(abstract:instance)]	1.89	1.833	-.624	.072
[T-S member:set]	0.78	1.093	.617	.077
[S-T elab(member:set)]	1.11	0.782	.646	.060
Qualitative gain				
Abstraction level				
T-S spec:gen	4.44	3.005	.855	.003**
[spec:gen]	16.00	5.050	.600	.088
Discourse relations				
situation:condition	8.89	2.667	.676	.045
[S-T elab(abstract:instance)]	1.89	1.833	-.661	.053
[S-T elab(step:process)]	2.22	1.481	.594	.092
Quantitative gain				
Abstraction level				
spec	3.33	2.550	-.699	.036
[T-S gen:spec]	31.33	11.424	-.662	.052
Discourse relations				
T-S object:attribute-direction	3.33	2.121	-.672	.047
S-T elab(set:member)	0.11	0.333	.884	.002**
S-T elab(whole:part)	1.44	1.590	.741	.022
S-T elab(gen:spec)	1.78	2.279	.680	.044
[object:attribute-direction]	4.56	2.789	-.595	.091
[situation:condition]	8.89	2.667	-.657	.055

Note. Trends are indicated by brackets. T = tutor; S = student; gen = general; spec = specific; elab = elaboration.

** $p < .01$.

ever, tutor specifications relative to the tutor's turn negatively correlated with overall gain. Consistent with the findings from the set of students taken together, one discourse relation whose frequency predicted overall gain among low pretesters was S-T situation:condition, in which the student asks the tutor to explain the circumstances under which a given physical state (velocity decreasing in the y direction) applies:

Example 13

Student: Why is velocity decreasing in the y direction?

Tutor: It starts out going up and gravity pulls it down. When acceleration is opposed to velocity, the object slows down.

Qualitative gain. Low pretesters' abstraction over the tutors' turns (T-S spec:gen) predicted qualitative gain score, consistent with a trend for abstraction either by the student or the tutor (spec:gen) to predict qualitative gain. Only one aggregate discourse relation variable significantly predicted qualitative gain among low pretest students: situation:condition, which is the conditional relation in which the second speaker provides the conditions that explain the situation described by the first speaker, either because the first speaker solicited this information or the second speaker initiated it. Example 13 illustrated a student-solicited conditional relation. The following exchange shows a tutor prompting the student to specify a condition in a T-S situation: condition relation:

Example 14

Tutor: Why does the tension equal the weight in this problem?

Student: Because there are no other outside forces acting on the bungee/jumper system.

Encouraging low pretest students to explain their claims (e.g., tension = weight) appears to be beneficial and is under the tutoring system's control, in contrast to student-initiated conditionals, such as the one shown in Example 13.

Quantitative gain. Consistent with the findings for all students considered together, the frequency of exchanges in which both participants focused on the case at hand negatively correlated with quantitative gain among low pretest students. In addition, the frequency of one type of specification negatively predicted quantitative gain for this group: T-S object:attribute-direction relations, in which the tutor prompts the student to specify the direction of a value. Specifying the correct direction of a vector often requires conceptual understanding, so this negative correlation could reflect the difficulty that less-prepared students have in determining direction. However, the frequency of several other specification relations predicted quantitative gains for low pretesters—in particular, tutor-initiated set:member [S-T elab(set:member)], whole:part [S-T elab(whole:part)], and gen:spec [S-T elab(gen:spec)] relations. The following exchange illustrates the tutor adding more specific information to the student's dialogue turn, in an S-T elab(gen:spec) relation:

Example 15

Reflection question: Does gravity have any effect on the vertical motion of the firecracker? What about the horizontal motion? Explain your answers.

Student: Vertical motion, yes; it makes it harder for the firecracker to travel away from the earth because gravity is pushing down, so it adds resistance.

Tutor: Good, that is right (and it pulls the firecracker back down after the high point also).

As this exchange illustrates, students sometimes answer questions correctly but not completely. The tutor added information necessary to complete the student's answer to the reflection question. Perhaps making low pretest students aware of complete answers, by adding to students' dialogue contributions, increases these students' quantitative problem-solving ability.

Discourse Relations That Predict Learning Among High Pretest Students

Correlations for high pretest students ($N = 7$) are displayed in Table 7. We again focus our discussion on significant findings ($p \leq .05$) for all three types of gain.

Overall gain. The frequency of only one discourse relation significantly predicted high pretest students' overall gain score: S-T elab(whole:part), which was also observed for the group of students as a whole. As discussed previously, this relation typically occurs when the tutor specifies the components of a vector that the student named, the specific forces that comprise the net force, etc. This finding suggests that adding this level of precision to high pretesters' dialogue contributions supports learning.

Table 7
Correlations for High Pretest Students (N = 7)

Abstraction level and discourse relations	Mean	SD	R	p
Overall gain				
Abstraction level: [S-T Gen]	3.00	1.826	-.700	.080
Discourse relations				
S-T elab(whole:part)	0.43	0.787	.826	.022
[object:attribute-units]	1.14	0.690	.709	.074
[object:attribute-direction]	3.43	1.813	-.713	.072
[all-object-attribute]	5.43	1.272	-.695	.083
[S-T elab(member:set)]	0.71	0.488	.719	.069
Qualitative gain				
Abstraction level				
[T-S spec]	2.29	1.704	.723	.066
[T-S spec:gen]	4.86	2.734	-.733	.061
[Gen]	10.43	5.224	-.688	.087
Discourse relations				
S-T elab(term:definition)	0.29	0.756	.863	.012
object:attribute-units	1.14	0.690	.817	.025
S-T whole:part	0.14	0.378	.863	.012
S-T elab(whole:part)	0.43	0.787	.809	.028
T-S elab(condition:situation)	0.29	0.756	.863	.012
[situation:condition]	9.57	3.359	-.697	.082
Quantitative gain				
Abstraction level: [Gen:spec]	32.00	11.986	-.720	.068
Discourse relations				
T-S step:process	1.00	1.000	.831	.020
S-T elab	18.00	6.325	-.756	.049
all-elab-bd	19.00	6.952	-.762	.046
[S-T elab(instance:abstract)]	2.57	1.397	-.733	.061
[all-abstract:instance-bd]	11.71	5.707	-.739	.058
[step:process]	1.71	1.113	.694	.084

Note. Trends are indicated by brackets. S = student; T = tutor; Gen = general; elab = elaboration; bd = bidirectional.

Qualitative gain. The frequency of several discourse relations predicted qualitative gains among high pretest students: tutor definitions of terms mentioned in the student's dialogue move [S-T elab(term:definition)]; whole:part relations [S-T whole:part, and S-T elab(whole:part)]; conditional relations that the student takes the initiative to complete [T-S elab(condition:situation)]; and one aggregate variable—object:attribute-units, in which the tutor prompts the student to provide missing units, or does this for the student. Tutor-initiated definitions typically occurred when the student used a term incorrectly and the tutor corrected it, as illustrated in the following exchange:

Example 16

Student: The force equals the mass of the book plus the other forces acting on it, which would be considered the acceleration.

Tutor: Well . . . the acceleration is the rate of change of its velocity.

Perhaps giving high pretest students the definition of a misused term sometimes suffices to correct their knowledge.

It is unclear why providing units (or prompting students to provide units) might support qualitative understanding. Perhaps units cement the difference between concepts or support students in understanding the temporal and spatial properties of physical concepts.

Quantitative gain. The frequency of one discourse relation predicted quantitative learning among high pretest students: ex-

changes in which the tutor provides a step in a line of reasoning and prompts the student to provide the line of reasoning that follows from that step or that is necessary to get to that step. For example, in the following exchange, the tutor states the final step in the problem (tension = weight) and prompts the student to explain how she arrived at that conclusion, via a T-S step:process relation:

Example 17

Tutor: OK . . . so then why does tension = weight . . . show me how you got your answer.

Student: $F = F_{\text{ten}} - ma$, $a = 0$, so $mg = F_{\text{ten}}$.

Two aggregate variables indicate that elaborations potentially hinder high pretesters' ability to gain quantitative knowledge and skills: S - T elab and all-elab-bd. The latter includes all elaborations initiated by either students or tutors; however, most were issued by tutors (354 vs. 16). Perhaps filling in too many details in the line of reasoning hinders learning among more knowledgeable students; it might be better to let them fill in the gaps on their own, as indicated by prior research on textual coherence (e.g., McNamara, 2001; McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996).

Decision Rules to Guide Automated Scaffolding

The analyses discussed in the previous section suggest that particular forms of cooperative execution that take place during scaffolding, implemented via co-constructed discourse relations, predict learning gains. However, since correlation does not imply causality, we need to determine if a tutorial dialogue system that is explicitly designed to encourage joint construction of these potentially beneficial discourse relations outperforms a counterpart tutoring system not so designed.

This section describes decision rules that stem from the findings discussed in the preceding section. These rules can guide the tutoring system in simulating these potentially effective aspects of human tutoring. Where appropriate, we provide further detail on the context in which these rules apply than we did in the previous section. In the next section, we illustrate how these decision rules are implemented in Rimac, in contrast to a control dialogue system.

Rule 1. *When the student provides a step in a line of reasoning, the tutor may provide the missing steps of the line of reasoning, rather than ask about each step individually.*

This decision rule stems from several correlations involving the step:process relation—specifically, for the group of students taken as a whole, the frequency of S-T elab(step:process) relations predicted overall gain, $R(14) = .646$, $p = .007$, and the aggregate variable step:process predicted both overall gain and qualitative gain, $R(14) = .582$, $p = .18$, and $R(14) = .591$, $p = .016$, respectively. The tutor's extension of the student's line of reasoning took place in three main contexts: (a) when the student answered a question correctly but not completely, as illustrated in Example 10; (b) when the student had some trouble coming up with a problem-solving or reasoning step, in which case the tutor filled in some of the line of reasoning and then prompted the student for additional steps; and (c) when the student reached the final step of a solution or line of reasoning,

in which case the tutor summarized the steps leading up to that conclusion. This mainly happened at the end of a problem.

Rule 2. *If a student states a value but does not state how he derived it, the tutor should prompt the student to explicate his reasoning process.*

This rule is similar to the preceding one, except that here the student, not the tutor, is expanding the line of reasoning as illustrated in Example 17. It stems from the finding that the frequency of T-S step:process relations predicted quantitative learning gains, particularly for high pretest students, $R(5) = .831$, $p = .020$.

Rule 3. *When students state vectors rather than vector components while solving equations, the tutor should provide the corresponding equation with components. Alternatively, the tutor should prompt the student to provide the vector components.*

This rule stems from several correlations involving the basic whole:part relation. For example, the frequency of S-T elab(whole:part) relations, in which the tutor specifies the vector components (Example 12), predicted overall gain for the whole group of students, $R(14) = .524$, $p = .037$. In addition, two aggregate variables predicted overall gain: whole:part and all-whole:part-bd, $R(14) = .528$, $p = .035$, and $R(14) = .553$, $p = .026$, respectively. Similar correlations were found for the group of high pretest students.

Rule 4. *When the student oversimplifies the circumstances under which a given physical situation applies or fails to make explicit the relationship between a narrower term and a broader term, the tutor should make these "member:set" relations explicit.*

This rule is based on the finding that the frequency of S-T elab(member:set) relations predicted overall gain for all students taken together, $R(14) = .667$, $p = .005$, for low pretest students, $R(7) = .646$, $p = .060$, and for high pretest students, $R(5) = 0.719$, $p = .069$. Example 11 illustrates a case in which the tutor states the class in which a narrower concept belongs (e.g., gravity is a type of acceleration) when the student's claim implies this but does not say it explicitly.

The following exchange illustrates the tutor reacting to a student's oversimplification of the circumstances associated with a physical situation. The student provides two examples of forces that could account for constant velocity (or a null net force); the tutor names the set "Anything else [other forces] that could make the net force 0":

Example 18

Student: No acceleration for a constant velocity; this would only be possible for a situation with a great deal of air resistance or friction.

Tutor: Or anything else to make the net force 0! The forces could be different.

Rule 5. *The tutor should ask "why" questions when the student does not provide an explanation to support a claim, especially with less knowledgeable students.*

This rule stems mainly from our finding that the frequency of conditional relations in which the tutor specified the conditions under which a situation described by the student applied (i.e., S-T situation:condition relations), correlated with overall learning gains for the group of low pretest students, $R(7) = .679$, $p = .044$. The aggregate variable situation:condition also predicted qualita-

tive gains for this group, $R(7) = .676$, $p = .045$. This finding supports Louwerse et al.'s (2008) suggestion that prompting students to express conditional relations exposes gaps in their reasoning process that the tutor can address, and this exercise promotes learning.

Example 13 illustrates a case in which a student takes initiative and asks the tutor to state the conditions that explain a given situation, while Example 14 illustrates the more readily implemented case of the tutor prompting the student to state relevant conditions, via a T-S situation:condition relation. "Why" prompts such as this typically occur when the student answers a question correctly but does not justify his answer, as in the following exchange:

Example 19

Reflection question: Does average acceleration imply that the acceleration is the same at every instant?

Student: No.

Tutor: Correct—could you say why?

Student: Because average is taking different velocities over different times.

Rule 6. *If the student answers a question incorrectly, if possible show why it is incorrect by stating the conditions under which it would be correct.*

This rule is related to the preceding and is mainly motivated by the correlation between the frequency of the aggregate situation:circumstance relation and qualitative gains among low pretest students. It reflects cases in which a student states a situation (the consequent in a conditional relation) and the tutor provides the conditions (antecedent) that would hold true if the situation were true. For example, in the following dialogue excerpt, the tutor states the conditions that would explain a net force of 0 on a bungee jumper:

Example 20

Reflection question: What minimum acceleration (in magnitude) must the jumper have in order for the cord not to break while he is on his way down?

Student: $700 \text{ N/mass} = a$.

Tutor: Not quite, good start. What is the "net" force on him? (in terms of the tension and mg)?

Student: The net force is 0.

Tutor: Ah, OK. When he is hanging there, it is 0, or if he is moving with constant velocity.

Rule 7. *If the student gives a partially correct answer, the tutor should complete it, especially for less knowledgeable students.*

This rule is based on the finding that the frequency in which the tutor extends a partial or underspecified statement in the student's dialogue turn, via S-T elab(gen:spec) relations, correlated with quantitative gains, among low pretest students. Example 15 demonstrates a tutor's application of this rule.

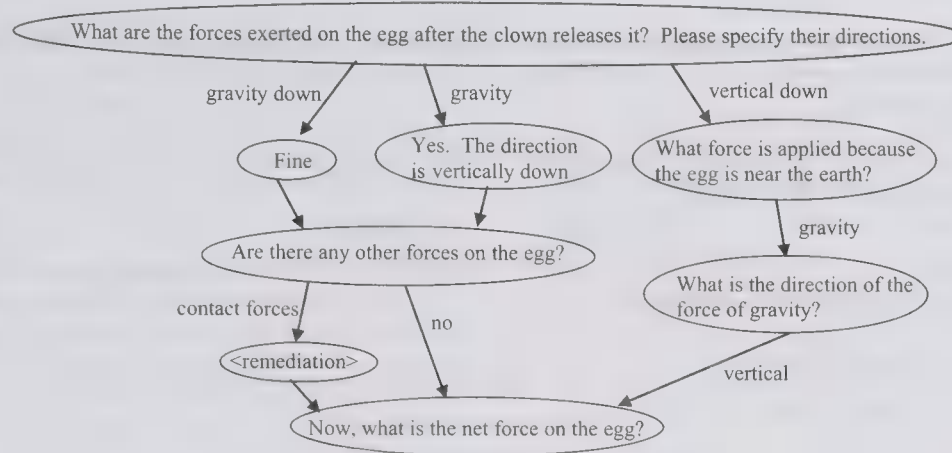


Figure 1. The dialogue paths of three students as they traverse the arcs in a knowledge construction dialogue (KCD). Adapted from "Tools for Authoring a Dialogue Agent That Participates in Learning Studies," by P. W. Jordan, B. Hall, M. Ringenberg, Y. Cui, and C. P. Rosé in R. Lucklin, K. R. Koedinger, & J. E. Greer (Eds.), 2007. *AIED 2007: Proceedings of the 13th International Conference on Artificial Intelligence in Education, Los Angeles, CA* (p. 48). Copyright 2007 by IOS Press, Amsterdam, the Netherlands Adapted with permission.

Rule 8. When the student uses a term incorrectly, give the definition of the term to help the student correct his or her mistake.

This rule stems from the finding that the frequency of S-T elab(term:definition) relations, in which the tutor defined a term that the student stated incorrectly or misapplied, correlated with qualitative gains, particularly among high pretest students, $R(5) = .863$, $p = .012$. Example 16 illustrates this rule.

Rule 9. The tutor should ask for missing units or prompt the student to provide them, especially when a student is performing well—for example, when the student is close to solving a problem or answering a qualitative question.

This rule is based on the finding that the frequency of the aggregate variable object:attribute-units, which includes all exchanges in which the student presented a value without units and the tutor either provided these units or prompted the student to do so, correlated with qualitative learning among high pretest students, $R(5) = .817$, $p = .025$. In the following exchange, the tutor provides the missing units:

Example 21

Student: $T - mg = ma$; $500 - 539 = 55a$.

Tutor: Good deal. (I would add units there by the way: $500N - 539N = 55 \text{ kg} \cdot a$.)

This rule is supported by prior research which used automated, machine learning methods to determine when abstractions and specification take place during reflective dialogues (Lipschultz, Litman, Jordon, & Katz, 2011). This research found that tutors tend to abstract over the student's dialogue contribution early in a reflective dialogue, when students are having difficulty responding to the tutoring system's reflection question. These abstractions appear to be aimed at ensuring that the student understands the basic concepts needed to answer the automated tutor's question. Then, as the dialogue progresses, and the student is closer to answering the reflection question correctly, specification becomes more frequent than abstraction, as tutors probe students for precision—for example, to

specify units and direction for a vector quantity, when the student only states its magnitude.

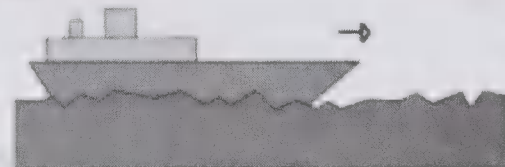
An Example: Implementing Decision Rules in Rimac

Reflective dialogues in Rimac were developed using a well-known framework called *knowledge construction dialogues* (KCDs). Several NL tutoring systems that have been found to be more effective than noninteractive forms of instruction (e.g., reading a textbook) have been developed using the KCD framework (e.g., Aleven, Popescu, Ogan, & Koedinger, 2003; Connelly & Katz, 2009; Graesser et al., 2003; Graesser, Person, Harter, & the Tutoring Research Group, 2001; Jordan, Ringenberg, & Hall,

Table 8

Example of a Rimac Problem and Reflection Question

A modern supertanker is gigantic: 1,200–1,300 ft. long with a 200-ft. beam. Fully loaded, it chugs along at about 16.0 knots (8.33 m/s). It can take 1,200 s to bring it to a full stop. Calculate both the acceleration in m/s^2 and the displacement of the tanker using this stopping time.



Acceleration:
Displacement:

Reflection question: Is the acceleration pointing in the same direction or in the opposite direction of the supertanker's motion?

Note. This example problem is part of the Andes Physics Tutor system, which was developed at Arizona State University and the University of Pittsburgh with support from the Pittsburgh Science of Learning Center, National Science Foundation Award SBE-0836012, and Office of Naval Research Grant N00014-96-1-0260 and is available at <http://www.andestutor.org>

Table 9
Application of Rule 5 in an Experimental Knowledge Construction Dialogue (KCD) but Not in the Corresponding Control KCD

Experimental KCD	Control KCD
<i>Tutor:</i> In the problem you’ve just solved, is the acceleration pointing in the same direction or in the opposite direction of the supertanker’s motion? <i>Student:</i> Opposite. <i>Tutor:</i> So why did you think this? <i>Student:</i> Because when an object is slowing down, acceleration is opposite to its motion. <i>Tutor:</i> Very good. Now let’s step through the reasoning.	<i>Tutor:</i> In the problem you’ve just solved, is the acceleration pointing in the same direction or in the opposite direction of the supertanker’s motion? <i>Student:</i> Opposite. <i>Tutor:</i> Very good. Now let’s step through the reasoning.

Note. Bold typeface indicates application of the rule.

2006; Katz et al., 2007; Rosé et al., 2001). During a KCD, the automated tutor presents a series of carefully ordered questions to the student, known as a *directed line of reasoning* (DLR; Evens & Michael, 2006). If a student answers a question correctly, the student advances to the next question in the DLR. Otherwise, the system launches a remedial subdialogue and then returns to the main DLR after the remedial subdialogue has completed. This process is illustrated in Figure 1.

KCDs in Rimac were implemented using TuTalk, a NL-dialogue-authoring toolkit (Jordan et al., 2006; Jordan, Hall, Ringenberg, Cui, & Rosé, 2007). TuTalk enables domain experts to construct NL tutoring systems without programming. Instead, they can focus on defining the tutoring content and structure of KCDs.

From a research perspective, the main advantage of using KCDs is that the content and structure of KCDs are determined a priori by the dialogue developer, so different versions of a given KCD can be designed to test a hypothesis. Since our goal was to determine if the decision rules that we specified to guide simulation of cooperative execution during scaffolding enhance learning, we developed two versions of each Rimac KCD: one version that implements these rules in appropriate contexts and another that simulates the standard KCD practice of the tutor eliciting information from the student, hinting when possible, and stating the answer after the student has made one or two unsuccessful tries.

We illustrate these two versions of a Rimac KCD with respect to the problem and reflection question shown in Table 8. Dialogue excerpts illustrate implementation of three of the decision rules described in the preceding section in the experimental version of the dialogue.

In the dialogue excerpt shown in Table 9, the decision-rule-driven KCD applies Rule 5 because the student answered the question correctly but without justifying it: *The tutor should ask “why” questions when the student does not provide an explanation to support a claim, especially for less knowledgeable students.* In contrast, the standard KCD excerpt just gives the student positive feedback. In both cases, the dialogue steps through the line of reasoning that leads to the correct answer because even students who answer “why” questions correctly sometimes do so by rote.

An intermediate step of the line of reasoning is shown in Table 10. In this case, the experimental (decision-rule-driven) KCD applies Rule 6: *If the student answers a question incorrectly, if possible show why it is incorrect by stating the conditions under which it would be correct.* In contrast, the standard KCD simply corrects the student without acknowledging or addressing the student’s response. Reformulating students’ dialogue contributions, whether correct or incorrect, is a common practice during human tutoring, but its effectiveness has not been directly tested. We implemented specific procedures within TuTalk to simulate reformulation in Rimac (Jordan, Katz, Albacete, Ford, & Wilson, 2012).

At the final step of the line of line of reasoning, the decision-rule-driven KCD summarizes the main physics principle addressed in this reflection question (see Table 8) and recaps the line of reasoning that supports this principle, thereby applying Rule 1. In contrast, the standard KCD only states the general principle, as shown in Table 11.

These alternative versions of Rimac are currently undergoing pilot evaluation during the kinematics and dynamics units in high school physics classes. Several urban and suburban schools in the Pittsburgh, PA, area are participating in this pilot study. We predict

Table 10
Application of Rule 6 in an Experimental Knowledge Construction Dialogue (KCD) but Not in the Corresponding Control KCD

Experimental KCD	Control KCD
<i>Tutor:</i> When an object is slowing down, how does the final velocity (v_f) compare to the initial velocity (v_i) for any interval of time? (smaller, larger, other) <i>Student:</i> Larger <i>Tutor:</i> If the object is speeding up, then its final velocity is larger than its initial velocity. But when an object is slowing down, its velocity is getting smaller all the time. So for any interval of time, the final velocity is smaller than the initial velocity.	<i>Tutor:</i> When an object is slowing down, how does the final velocity (v_f) compare to the initial velocity (v_i) for any interval of time? (smaller, larger, other) <i>Student:</i> Larger <i>Tutor:</i> When an object is slowing down, its velocity is getting smaller all the time. So for any interval of time, the final velocity is smaller than the initial velocity.

Note. Bold typeface indicates application of the rule.

Table 11
Application of Rule 1 in an Experimental Knowledge Construction Dialogue (KCD) but Not in the Corresponding Control KCD

Experimental KCD	Control KCD
<p><i>Tutor:</i> Hence from the definition of acceleration, how do the direction of the acceleration and the direction of the change in velocity relate? (same, opposite, no relation, it depends)</p> <p><i>Student:</i> They are the same.</p> <p><i>Tutor:</i> Very good. Hence, in the current problem, the direction of the acceleration opposes the direction of motion of the supertanker because the direction of the change in velocity opposes the motion of the supertanker. And this happens because the final velocity is smaller than the initial velocity, and velocity has the same direction as motion. Therefore, in general, we can say that when an object is slowing down, the direction of the acceleration opposes the motion of the object.</p>	<p><i>Tutor:</i> Hence from the definition of acceleration, how do the direction of the acceleration and the direction of the change in velocity relate? (same, opposite, no relation, it depends)</p> <p><i>Student:</i> They are the same.</p> <p><i>Tutor:</i> Very good. In general, when an object is slowing down, the direction of the acceleration opposes the motion of the object.</p>

Note. Bold typeface indicates application of the rule.

that the decision-rule-driven version will outperform the less interactive control and that the effect will be greatest for less prepared students, a common finding for evaluations of instructional interventions (VanLehn et al., 2007).

Conclusion

The holy grail of tutoring research is to identify specific features of human tutorial dialogue that account for its remarkable effectiveness (e.g., Bloom, 1984; Cohen, Kulik, & Kulik, 1982), so that these features can be simulated in NL tutoring systems. Although the interaction hypothesis posits that more interactive tutoring will result in more learning, research to test this hypothesis shows that constructs like interactivity and cooperative execution are too vague to guide automated tutoring and, in particular, the scaffolding that takes place when students are having difficulty solving a quantitative problem or answering a conceptual question. In order to operationalize interactivity and cooperative execution, we need to identify the linguistic mechanisms that implement these constructs during human one-on-one tutoring and determine which mechanisms enhance learning. This knowledge can then be used to formulate decision rules that can be implemented and tested within NL tutoring systems. The research described in this article takes a step in this direction.

Overall, this study supports the interaction hypothesis. Our analyses suggest that the effectiveness of human tutoring might very well lie in the language of tutoring itself—in particular, in the types of discourse relations that students and tutors co-construct during tutorial dialogues. Moreover, the types of co-constructed discourse relations that predict learning seem to vary according to students' ability levels. However, given the small sample size, these findings should be cross-validated by analyses of dialogue

corpora involving a larger number of subjects (both students and tutors).

A second limitation of this work stems from its focus on co-constructed discourse relations. It might well be the case that some discourse relations are better “told” than “elicited,” that is, conveyed through direct, didactic explanations, instead of co-constructed while questioning the student. For example, we were surprised that we did not find a relationship between the frequency with which a tutor stated abstract principles or formulae (e.g., the equation for Newton's second law) and prompted students to instantiate these principles, as captured by the T-S abstract:instance relation, and student learning. However, this does not negate the potential effectiveness of instantiation of variables, principles, and so on during tutoring. Perhaps the didactic form of this relation (abstract:instance) does support learning, among some groups of students, but our analyses did not investigate correlations between didactically delivered discourse relations and learning. Hence, one goal of our future work will be to compare the effectiveness of didactic and interactive forms of particular discourse relations.

A third limitation of this research is that we did not consider variations in the way that co-construction of discourse relations is carried out and how these variations might impact learning. For example, we observed that there are two main ways in which tutors address abstractions. Tutors either anchor discussions about concepts and principles in the case at hand (i.e., the current problem) or address these abstractions in context-independent terms. For example, in both dialogue excerpts shown in Table 12, the tutor addresses the conditional: *if an object travels upward and comes back down, its vertical displacement is 0*. In the excerpt shown in the left column, the tutor grounds this abstraction in the current

Table 12
Alternative Ways of Prompting for a Conditional Relation

Context-specific prompt for a conditional relation	Context-independent prompt to complete a conditional relation
<p><i>Tutor:</i> Picture in your mind's eye . . . firecracker goes up, and then comes down and lands on the ground. What is the net vertical displacement for that whole process?</p> <p><i>Student:</i> 0.</p>	<p><i>Tutor:</i> Regardless of whether we call ground level $y = 0$ or $y = 500$, what is the y component of the displacement for an object that goes up and then comes back down to ground level?</p> <p><i>Student:</i> 0 meters.</p>

physical situation about a firecracker. He provides the antecedent of the conditional (the “if clause”) and prompts the student for the consequent (the “then clause”). In contrast, in another dialogue about the same problem, shown in the right column of Table 12, the tutor speaks in more general, context-independent terms; he refers to “an object,” not to the firecracker. Future research should examine which approach (if either) is better and for which types of students.

One important lesson that automated approaches to identifying decision rules to guide tutoring has taught us is that the “right” pedagogical move in a given context can depend on many factors: student characteristics, features of the problem under discussion, features of the dialogue context, and so on. We might not even be able to specify the relevant factors a priori. It is quite likely that we find that the decision rules suggested by our analyses are underspecified and in need of refinement. Although most of these rules, as stated, could apply to any scientific, problem-solving domain, their generalizability remains to be tested. A combination of automated approaches and carefully controlled, experimental studies of “tuned” versions of these decision rules and others will bring tutoring researchers closer to cracking the code of interactivity and developing more effective tutoring systems as a result.

References

- Aleven, V., Popescu, O., Ogan, A., & Koedinger, K. (2003). A formative classroom evaluation of a tutorial dialogue system that supports self-explanation. In H. U. Hoppe, F. Verdejo, & J. Kay (Eds.), *AIED 2003: Proceedings of the 11th International Conference on Artificial Intelligence in Education, Sydney, Australia, July 2003* (pp. 39–46). Amsterdam, the Netherlands: IOS Press.
- Beck, J., Woolf, B., & Beal, C. (2000). ADVISOR: A machine learning architecture for intelligent tutor construction. In H. Kautz & B. Porter (Co-chairs), *Proceedings of the Seventeenth National Conference on Artificial Intelligence, Austin, TX* (pp. 552–557). Menlo Park, CA: AAAI Press.
- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13, 4–16.
- Boyer, K. E., Phillips, R., Ingram, A., Ha, E. Y., Wallis, M., Vouk, M., & Lester, J. (2010). Characterizing the effectiveness of tutorial dialogue with hidden Markov models. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems: Proceedings of the 10th international conference on intelligent tutoring systems, ITS 2010, Pittsburgh, PA, June 14–18, 2010* (Pt. 1, Lecture Notes in Computer Science 6094, pp. 55–64). Berlin, Germany: Springer-Verlag.
- Chi, M., Jordan, P., VanLehn, K., & Litman, D. (2009). To elicit or to tell: Does it matter? In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. Graesser (Eds.), *Proceedings of the 2009 Conference on Artificial Intelligence in Education. Building learning systems that care: From knowledge representation to affective modeling* (pp. 197–204). Amsterdam, the Netherlands: IOS Press.
- Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25, 471–533. doi:10.1207/s15516709cog2504_1
- Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2010). Inducting effective pedagogical strategies using learning context features. In P. DeBra, A. Kobsa & D. Chin (Eds.), *User Modeling, Adaptation and Personalization: 18th International Conference, UMAP 2010* (pp. 147–158). Heidelberg, Germany: Springer.
- Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011a). An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *International Journal of Artificial Intelligence in Education*, 21, 83–113.
- Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011b). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction*, 21, 137–180.
- Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13, 259–294. doi:10.1207/s15516709cog1302_7
- Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19, 237–248.
- Collins, A., & Brown, J. S. (1986). *The computer as a tool for learning through reflection* (Technical Rept. 376). Washington, DC: National Institute of Education.
- Connelly, J., & Katz, S. (2009). Towards more robust learning of physics via reflective dialogue extensions. In C. Fulford (Ed.), *ED-MEDIA 2009: World Conference on Educational Multimedia, Hypermedia, & Telecommunications, Honolulu, Hawaii, June 22–26, 2009*. Chesapeake, VA: Association for the Advancement of Computing in Education.
- Di Eugenio, B., & Green, N. L. (2010). Emerging applications of natural language generation in information visualization, education, and health care. In N. Indurkha & F. J. Damerau (Eds.), *Handbook of natural language processing* (2nd ed.; pp. 557–576). Boca Raton, FL: Chapman & Hall/CRC.
- Evens, M. W., & Michael, J. A. (2006). *One-on-one tutoring by humans and machines*. Mahwah, NJ: Erlbaum.
- Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, 8, 8–11. doi:10.1016/j.tics.2003.10.016
- Gick, M., & Holyoak, K. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1–38. doi:10.1016/0010-0285(83)90002-6
- Gick, M., & Holyoak, K. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of learning: Contemporary research and applications* (pp. 9–46). New York, NY: Academic Press.
- Graesser, A. C., Jackson, G. T., Mathews, E. C., Mitchell, H. H., Olney, A., Ventura, M., . . . Person, N. K. (2003). Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialogue. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society 2003* (pp. 474–479). Boston, MA: Cognitive Science Society.
- Graesser, A. C., Person, N. K., Harter, D., & the Tutoring Research Group. (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, 12, 257–279.
- Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology*, 9, 495–522. doi:10.1002/acp.2350090604
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English* (English Language Series). London, England: Pearson Education.
- Halloun, I., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, 53, 1043–1055. doi:10.1119/1.14030
- Hovy, E. (1990). Parsimonious and profligate approaches to the question of discourse structure relations. In K. R. McKeown, J. D. Moore, & S. Nirenburg (Eds.), *Proceedings of the Fifth International Workshop on Natural Language Generation, June 3–6, 1990, Dawson, PA* (pp. 59–65). Stroudsburg, PA: Association for Computational Linguistics Special Interest Group on Natural Language Generation (SIGGEN).
- Jordan, P. W., Hall, B., Ringenberg, M., Cui, Y., & Rosé, C. P. (2007). Tools for authoring a dialogue agent that participates in learning studies. In R. Lucklin, K. R. Koedinger, & J. E. Greer (Eds.), *AIED 2007: Proceedings of the 13th International Conference on Artificial Intelligence in Education, Los Angeles, CA* (pp. 43–50). Amsterdam, the Netherlands: IOS Press.
- Jordan, P. W., Katz, S., Albacete, P., Ford, M., & Wilson, C. (2012).

- Reformulating student contributions in tutorial dialogue. In B. Di Eugenio, S. McRoy, A. Gatt, A. Betz, A. Koller, & K. Striegnitz (Eds.), *INGL 2012: Proceedings of 7th International Natural Language Generation Conference, Utica, IL, May 30–June 1, 2012* (pp. 95–99). Stroudsburg, PA: Association for Computational Linguistics Special Interest Group on Natural Language Generation (SIGGEN).
- Jordan, P. W., Ringenber, M., & Hall, B. (2006). Rapidly developing dialogue systems that support learning studies. In E. Lulis & P. Wiemer-Hastings (Eds.), *Proceedings of ITS 2006 Workshop on Teaching with Robots, Agents, and NLP*. Jhongli, Taiwan: National Center University Research Center for Science and Technology for Learning.
- Katz, S., Allbritton, D., & Connelly, J. (2003). Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence and Education*, 13, 79–116.
- Katz, S., Connelly, J., & Wilson, C. (2007). Out of the lab and into the classroom: An evaluation of reflective dialogue in Andes. In R. Lucklin, K. R. Koedinger, & J. E. Greer (Eds.), *AIED 2007: Proceedings of the 13th International Conference on Artificial Intelligence in Education, Los Angeles, CA* (pp. 425–432). Amsterdam, the Netherlands: IOS Press.
- Lee, A., & Hutchison, L. Improving learning from examples through reflection. *Journal of Experimental Psychology: Applied*, 4, 187–210.
- Leher, R., & Littlefield, J. (1993). Relationships among cognitive components in logo learning and transfer. *Journal of Educational Psychology*, 85, 317–330. doi:10.1037/0022-0663.85.2.317
- Lipschultz, M., Litman, D., Jordan, P., & Katz, S. (2011). Predicting changes in level of abstraction in tutor responses to students. In R. C. Murray & R. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2011), May 18–20, 2011, Palm Beach, FL*. Miami, FL: FLAIRS.
- Litman, D., & Forbes-Riley, K. (2006). Correlations between dialogue acts and learning in spoken tutoring dialogues. *Natural Language Engineering*, 12, 161–176. doi:10.1017/S1351324906004165
- Louwerse, M. M., Crossley, S. A., & Jeuniaux, P. (2008). What if? Conditionals in educational registers. *Linguistics and Education*, 19, 56–69. doi:10.1016/j.linged.2008.01.001
- Mann, W. C., & Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8, 243–281. doi:10.1515/text.1.1988.8.3.243
- McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology*, 55, 51–62.
- McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, 22, 247–287. doi:10.1080/01638539609544975
- McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14, 1–43. doi:10.1207/s1532690xcil401_1
- Murray, R. C., & VanLehn, K. (2006). A comparison of decision-theoretic, fixed-policy and random tutorial action selection. In K. Ashley & M. Ikeda (Eds.), *Intelligent tutoring systems: Proceedings of the eighth international conference, ITS 2006, Jhongli, Taiwan, June 26–30, 2006* (Lecture Notes in Computer Science 4053, pp. 114–123). Berlin, Germany: Springer-Verlag. doi:10.1007/11774303_12
- Obama, B. (2009). *Remarks by the president at the National Academy of Sciences Annual Meeting*. Retrieved March 13, 2013, from http://www.whitehouse.gov/the_press_office/Remarks-by-the-President-at-the-National-Academy-of-Sciences-Annual-Meeting
- Pilkington, R. (2001). Analysing educational dialogue interaction: Towards models that support learning. *International Journal of Artificial Intelligence in Education*, 12, 1–7.
- Ravenscroft, A., & Pilkington, R. M. (2000). Investigation by design: Developing models to support reasoning and conceptual change. *International Journal of Artificial Intelligence in Education*, 11, 273–298.
- Reed, S. K. (1993). A schema-based theory of transfer. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, Cognition, and instruction* (pp. 39–67). Norwood, NJ: Ablex.
- Rosé, C., Jordan, P., Ringenber, M., Siler, S., VanLehn, K., & Weinstein, A. (2001). Interactive conceptual tutoring in Atlas-Andes. In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.), *Artificial intelligence in education* (pp. 256–266). Amsterdam, the Netherlands: IOS Press.
- Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanisms of a neglected phenomenon. *Educational Psychologist*, 24, 113–142. doi:10.1207/s15326985ep2402_1
- Steinhaus, N., Campbell, G. E., Taylor, L. S., Caine, S., Scott, C., Dzikovska, M., & Moore, J. D. (2011). Talk like an electrician: Mimicking behavior in an intelligent tutoring system. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education: 15th international conference, AIED 2011, Auckland, New Zealand, June 28–July 2, 2011* (Lecture Notes in Artificial Intelligence 6738, pp. 361–368). Berlin, Germany: Springer.
- Tchetagni, J. M. P., Nkambou, R., & Bourdeau, J. (2007). Explicit reflection in prolog-tutor. *International Journal of Artificial Intelligence in Education*, 17, 169–215.
- van de Sande, C., & Greeno, J. G. (2010). A framing of instructional explanations: Let us explain with you. In M. K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines* (pp. 69–82). New York, NY: Springer.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46, 197–221. doi:10.1080/00461520.2011.611369
- VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, 31, 3–62. doi:10.1080/03640210709336984
- VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., . . . Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, 15, 1–47.
- Ward, A., Connelly, J., Katz, S., Litman, D., & Wilson, C. (2009). Cohesion, semantics, and learning in reflective dialog. In S. D. Craig & D. Dicheva, *AIED 2009: 14th International Conference on Artificial Intelligence in Education Workshop Proceedings. Vol. 10: Natural Language Processing in Support of Learning. Metrics, Feedback, and Connectivity*. Available at <http://webu2.upmf-grenoble.fr/sciedu/nlppl>
- Ward, A., & Litman, D. (2008). Semantic cohesion and learning. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), *Intelligent tutoring systems: Proceedings of the 9th International Conference, ITS 2008, Montreal, Canada, June 23–27, 2008* (Lecture Notes in Computer Science 5091, pp. 459–469). New York, NY: Springer.
- Ward, A., & Litman, D. (2011). Adding abstractive reflection to a tutorial dialog system. In R. C. Murray & R. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2011), May 18–20, 2011, Palm Beach, FL* (Paper 2575). Miami, FL: FLAIRS.

Received December 15, 2011

Revision received October 22, 2012

Accepted December 18, 2012 ■

Human and Automated Assessment of Oral Reading Fluency

Daniel Bolaños, Ron A. Cole, and Wayne H. Ward
Boulder Language Technologies, Boulder, Colorado

Gerald A. Tindal
University of Oregon

Jan Hasbrouck
Gibson Hasbrouck & Associates, Wellesley, Massachusetts

Paula J. Schwanenflugel
The University of Georgia

This article describes a comprehensive approach to fully automated assessment of children's oral reading fluency (ORF), one of the most informative and frequently administered measures of children's reading ability. Speech recognition and machine learning techniques are described that model the 3 components of oral reading fluency: word accuracy, reading rate, and expressiveness. These techniques are integrated into a computer program that produces estimates of these components during a child's 1-min reading of a grade-level text. The ability of the program to produce accurate assessments was evaluated on a corpus of 783 one-min recordings of 313 students reading grade-leveled passages without assistance. Established standardized metrics of accuracy and rate (words correct per minute [WCPM]) and expressiveness (National Assessment of Educational Progress Expressiveness scale) were used to compare ORF estimates produced by expert human scorers and automatically generated ratings. Experimental results showed that the proposed techniques produced WCPM scores that were within 3–4 words of human scorers across students in different grade levels and schools. The results also showed that computer-generated ratings of expressive reading agreed with human raters better than the human raters agreed with each other. The results of the study indicate that computer-generated ORF assessments produce an accurate multidimensional estimate of children's oral reading ability that approaches agreement among human scorers. The implications of these results for future research and near term benefits to teachers and students are discussed.

Keywords: oral reading fluency, automated reading assessment, expressive reading, automatic speech recognition

Reading assessments provide school districts and teachers with critical and timely information for identifying students who need immediate help; for making decisions about reading instruction; for monitoring individual student's progress in response to instruc-

tional interventions; for comparing different approaches to reading instruction; and for reporting annual outcomes in classrooms, schools, school districts, and states. One of the most common tests administered to primary school students is *oral reading fluency* (ORF). Over 25 years of scientifically based reading research has established that fluency is a critical component of reading and that effective reading programs should include instruction in fluency (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Kuhn & Stahl, 2000; National Reading Panel, 2000). Although ORF does not measure comprehension directly, there is substantial evidence that estimates of ORF predict future reading performance and correlate strongly with comprehension (Fuchs et al., 2001; Shinn, 1998). According to Wayman, Wallace, Wiley, Tichá, and Espin (2007), ORF is a valid indicator of comprehension in early grades, though less so beyond Grade 4. Because ORF can be measured rather quickly (typically in 5–10 min) with good validity and reliability, it is widely used to screen individuals for reading problems and to measure reading progress over time.

In this article, we present a comprehensive approach to assessing ORF accurately and automatically through the use of speech recognition and machine learning techniques. The approach is comprehensive because all three measures of ORF—accuracy, rate (combined into a words correct per minute [WCPM] score), and expressiveness—can be measured automatically and in real time, whereas expressiveness is rarely scored in real-world educational

This article was published Online First September 9, 2013.

Daniel Bolaños, Ron A. Cole, and Wayne H. Ward, Boulder Language Technologies, Boulder, Colorado; Gerald A. Tindal, Department of Psychology, University of Oregon; Jan Hasbrouck, Gibson Hasbrouck & Associates, Wellesley, Massachusetts; Paula J. Schwanenflugel, Department of Psychology, The University of Georgia.

This work was supported by U.S. Department of Education Award Number R305B070434, National Science Foundation Award Number 0733323, and National Institutes of Health Award Number R44 HD055028. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute of Education Sciences or the National Science Foundation. We gratefully acknowledge the help of Angel Stobaugh, director of literacy education at Boulder Valley School District, and the principals and teachers who allowed us to visit their schools and classrooms. We appreciate the amazing efforts of Jennifer Borum and the efforts of Linda Hill, Suzan Heglin, and the rest of the human experts who scored the text passages.

Correspondence concerning this article should be addressed to Daniel Bolaños, Boulder Language Technologies, 2960 Center Green Court, Boulder, CO 80301. E-mail: dani@bltek.com

contexts. The ultimate goal of research leading to fully automatic and comprehensive assessment of ORF is to provide an accurate, accessible, and low-cost alternative to human-administered assessments. Successful outcomes of research in this area would substantially reduce the millions of hours that teachers spend each year assessing their students' reading abilities, which is mandated by federal law in the United States. In addition, computer-based assessments of ORF could generate detailed records of individual student's performance, including the digital recordings of each reading session that could be reviewed by teachers, parents, and students, and analyzed automatically for detailed information about the student's reading problems. Automatic administration of ORF will also enable collection of massive amounts of speech data that can be used to analyze and understand children's development of reading skills; these data can also be used to improve the performance of the speech recognition technologies.

We used a speech recognition system (Bolaños, 2012) specifically designed to process children's read speech to produce a word-level hypothesis of what the student read from a grade-level text during 1 min. From this hypothesis and the text passage, a WCPM score was computed reflecting the student's reading accuracy and rate. In order to assess prosodic reading, we developed a series of lexical and prosodic features that were extracted from the student's speech. These included analysis of the text syntax and its correlation with filled pauses and silence regions, syllable and word duration, pitch, and word co-occurrences, among other features described below. Machine learning classifiers were trained on these features, resulting in statistical models that were able to discriminate between different degrees of prosodic reading using the National Assessment of Educational Progress ORF Scale (NAEP; Daane, Campbell, Grigg, Goodman, & Oranje, 2005). A hierarchical classification scheme was used in order to assign 1-min reading sessions to levels in the NAEP scale.

The accuracy of these assessment methods was evaluated on approximately 13 hr of speech collected from the 313 first-through fourth-grade students who read grade-level text passages. WCPM scores as well as NAEP assessments generated by the system, FLuent Oral Reading Assessment (FLORA), were compared with those produced by at least two independent human judges.

The remainder of the article is organized as follows: The next section provides the scientific rationale for assessing ORF. We then describe the corpus of children's read speech that was collected for this study. We then describe the system and features used to assess WCPM (accuracy and rate) and expressive reading using lexical and prosodic features extracted from the speech. The last section presents the discussion and conclusions.

Scientific Rationale for FLORA

ORF

ORF is typically defined as a student's ability to read words in grade-level texts accurately and effortlessly, at a natural speech rate and with appropriate prosodic expression. A synthesis of scientifically based reading research by the National Reading Panel (2000) concluded that

Reading fluency is one of several critical factors necessary for reading comprehension, but it is often neglected in the classroom. If children

read out loud with speed, accuracy and proper expression, they are more likely to comprehend and remember the material than if they read with difficulty and in an inefficient way.

Accuracy and automaticity. Accurate reading speed is both a strong discriminator of reading ability (e.g., Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003; Perfetti, 1985) and a strong predictor of later reading proficiency (Lesgold & Resnick, 1982; Scarborough, 1998; see review by Compton & Carlisle, 1994.) As Jenkins et al. (2003) put it: "Together with listening comprehension, word-reading skill accounts for nearly all of the reliable variance in reading ability, and individual differences in word recognition explain significant variance in reading ability, even after controlling for reading comprehension" (Curtis, 1980; Hoover & Gough, 1990).

ORF depends on the ability to recognize words in a text *quickly and automatically*. As defined by Fuchs et al. (2001), automaticity is "the oral translation of text with speed and accuracy." Automaticity theory (LaBerge & Samuels, 1974; Samuels, 1985; Wolf, 1999) and related verbal efficiency accounts of reading (Perfetti, 1985) hold that students who have learned to decode printed words automatically are able to devote more attention (cognitive resources) to comprehending what they are reading. Readers who have not achieved automaticity during word recognition must devote significant attention to recognizing words (at the expense of devoting this attention to making sense of the text), resulting in slower reading times and weaker comprehension. Support for automaticity and verbal efficiency theories of reading is provided by the strong association between the speed of reading words, either in word lists or in context, and measures of reading comprehension.

Expressiveness. Although readers who have achieved fluency can read texts rapidly and accurately, they may not read expressively (i.e., they may not pause between sentences, at major phrase boundaries within sentences, or produce appropriate prosody when reading out loud). Expressive reading is the third critical component of *reading fluency*, typically defined as reading a text with the appropriate expression, intonation, and phrasing in order to preserve meaning (Miller & Schwanenflugel, 2008).

Connection between ORF and comprehension. For over 25 years, researchers have documented the association between reading fluency and comprehension. Reviews of the research on ORF have demonstrated consistently moderate to strong correlations between ORF and comprehension (Marston, 1989; Shinn, 1998). Research results have demonstrated high concurrent validity between ORF and measures of word recognition and reading comprehension (Hosp & Fuchs, 2005; Jenkins et al., 2003), and between ORF and nationally normed standardized tests of reading comprehension (Roehrig, Petscher, Nettles, Hudson, & Torgesen, 2008; Schilling, Carlisle, Scott, & Zeng, 2007; Schwanenflugel et al., 2006). Measures of ORF in early grades have also been found to predict comprehension in later grades (Kim, Petscher, Schatschneider, & Foorman, 2010). Thus, the relation between ORF and reading comprehension has been well established by previous research, particularly for students in elementary school (Kim et al., 2010; Roberts, Good, & Corcoran, 2005; Roehrig et al., 2008).

Previous Work Using Automatic Speech Recognition to Assess and Improve ORF

Automatic assessment of reading accuracy and rate. Over two decades of research has investigated the use of automatic speech recognition (ASR) to assess and improve reading. Seminal research conducted by Jack Mostow and his colleagues in Project Listen at Carnegie Mellon University has demonstrated the effectiveness of ASR for improving reading fluency and comprehension for both native and nonnative speakers of English (Mostow et al., 2003; Reeder, Shapiro, & Wakefield, 2007). Mostow et al. (2003) used an ASR system to measure a student's *interword latency*, defined as the elapsed time between certain words read aloud by the student that were scored as correctly read by the ASR system. Their model of interword latency produced a correlation of over .7 with independent WCPM measures of ORF using grade-level passages.

In the context of Project Tball (Technology Based Assessment of Language and Literacy) at the University of California, Los Angeles and University of Southern California, Black, Tepperman, Lee, and Narayanan (2008) investigated oral reading of 55 isolated words produced by kindergarten, first-, and second-grade children with the aim of detecting reading miscues automatically, such as sounding-out, hesitations, whispering, elongated onsets, and question intonations. Black et al. developed an ASR system that used specialized grammars to model word-level disfluencies using the subword-modeling approach developed by Hagen and Pellom (2005). Scores produced by the recognition system correlated highly (.91) with fluency judgments provided by human listeners.

A series of studies by Bryan Pellom and Andreas Hagen and their collaborators (Hagen, Pellom, & Cole, 2007) investigated ways to optimize an ASR system for children's read speech. The research resulted in a reduction in the word error rate from 17.4% to 7.6%. Hagen et al. (2007) developed a version of the ASR system that used subword-modeling rather than whole-word scoring to detect reading errors. In the study, several subword lexical units and approaches were evaluated for detection of reading disfluencies, and modest gains were reported. Bolaños (2008) reported that additional detection gains were achieved by using syllable graphs to represent hypotheses from the ASR system.

Automatic assessment of expressive oral reading. Although the National Reading Panel (2000) and research community define ORF in terms of word recognition accuracy, reading rate, and how expressively the student reads (see Kuhn, Schwanenflugel, & Meisinger, 2010, for a discussion of this topic), expressiveness is rarely measured in assessments of ORF. Only recently has the expressiveness aspect of the reading fluency construct found its way into automated assessments of fluency. Duong, Mostow, and Sitaram (2011) investigated two alternative methods of measuring prosody during children's oral reading. The first method, which was text dependent, consisted of generating a prosodic template model for each sentence in the text. The template was based on word-level features like pitch, intensity, latency, and duration extracted from fluent adult narrations. The second method investigated adult narrations to train a general duration model that could be used to generate expected prosodic contours of sentences for any text, so an adult reader was no longer required to generate sentence templates for each new text. Both methods were evaluated for their ability to predict student's scores on fluency and

comprehension tests, and each produced promising results, with the second, automated method for generating prosodic sentence templates outperforming the system that compared children's read speech with adult narrations of each individual sentence in the text. However, neither of these methods could satisfactorily classify sentences using the NAEP expressiveness rubric relative to human judgments, which was probably due to the low human interrater reliability reported in this study.

Development of the FLORA System

Development of a Corpus for Assessing ORF

Data collection setting. Data were collected from 313 first-through fourth-grade students in four elementary schools (nine classrooms) in the Boulder Valley School District in Colorado. Data were collected from students in their classrooms at their schools. School 1 had 53.8% students receiving free or reduced lunches, and the lowest literacy achievement scores of the three schools on the Colorado state literacy test given to third-grade students; 53% third-grade students in School 1 scored proficient or above on the state reading assessment. School 2 had 51.7% students with free or reduced lunch (similar to School 1), but 79% of third-grade students tested as proficient or above on the state literacy test. School 2 was a bilingual school with nearly 100% English learners (ELs) who spoke Spanish as their first language. School 3 had 18.4% of students with free or reduced lunch, 85% of students were proficient or above in the state literacy test. School 3 also had relatively few ELs.

Text passages. Twenty text passages were available for reading at each grade level. The standardized text passages were downloaded from a website (Good & Kaminski, 2002) and are freely available for noncommercial use. The text passages were designed specifically to assess ORF and are about the same level of difficulty within each grade level. ORF norms have been collected for these text passages for tens of thousands of students at each grade level in fall, winter, and spring semesters, so that students can be assigned to percentiles based on national WCPM scores (Hasbrouck & Tindal, 2006).

Data collection protocol. The data were collected using the FLORA system (Bolaños, Cole, Ward, Borts, & Svirsky, 2011), which was configured to enroll each student, randomly select one passage from the set of 20 standardized passages for the student's grade level, and present the passage to the student for reading out loud. Because testing was conducted in May, near the end of the school year, classroom teachers had recently assessed their student's oral reading performance (using text passages different from those used in our study). About 20% of the time, teachers requested that specific students be presented with text passages either one or two levels below or one or two levels above the student's grade level. Thus, about 80% of students in each grade read passages at their grade level, whereas 20% of students read passages above or below their grade level, based on their teachers' recommendations. Depending on the number of students who needed to be tested on a given day, each student was presented with two or three text passages to read aloud.

During the testing procedure, the student was seated before a laptop and wore a set of headphones with an attached noise-cancelling microphone. The experimenter observed or helped the

student enroll in the session, which involved entering the student's gender, age, and grade level. FLORA then presented a text passage, started the 1-min recording at the instant the passage was displayed, recorded the student's speech, and relayed the speech to a server.

Corpus summary. The corpus comprised 783 recordings from 313 first- through fourth-grade students for a total of approximately 13 hr of speech data. Each recording was scored manually by two human judges. Words were scored as reading errors if the word was skipped over, or the judge decided that the word was misread. Insertions of words (intrusions) were not scored as reading errors, as insertions were not counted as errors in the national norms collected by Hasbrouck and Tindal (Hasbrouck & Tindal, 2006).

Automatic Generation of WCPM Scores

The number of words that a student read correctly during 1 min was computed automatically by ReadToMe, the reading tracker built on top of our ASR system (Bolaños, 2012). The computation of the WCPM score was done as follows. (a) ReadToMe used the Baviaca speech recognition toolkit (Bolaños, 2012) to produce a word-level hypothesis representing what the student read. (b) ReadToMe aligned the hypothesis to the reference text (the words in the text passage the student read) and tagged each of the words in the reference text as correctly or incorrectly read or skipped over. (c) Finally, ReadToMe counted the number of words scored as correctly read during the 1-min reading; this number is the resulting WCPM score for the text passage.

Automatic Assessment of Expressive Reading

In order to assess expressive reading automatically, we proposed a set of lexical and prosodic features that can be used to train a machine learning system to classify how expressively students read text passages aloud using the 4-point NAEP scale. The proposed features were designed to measure the speech behaviors associated with each of the four levels of fluency described in the NAEP rubric and were informed by research on acoustic-phonetic, lexical, and prosodic correlates of fluent and expressive reading described in the research literature (Kuhn et al., 2010). Features were extracted from multiple sources, including the recognition hypothesis, a pitch-extractor, and a syllabification tool. Features included the WCPM score itself, the speaking rate, sentence reading rate, number of word repetitions, location of the pitch accent, word and syllable durations, and filled and unfilled pauses and their correlation to punctuation marks in the text passage. A detailed description, motivation, and analysis of all the features proposed and used for the study can be found in Bolaños et al. (2013).

Classification method. In order to classify the 783 one-min recordings using the features proposed, we used a powerful classification technique called support vector machines (Vapnick, 1995). We experimented with difference classification strategies and found a strategy based on a decision directed acyclic graph (DAG) to be most successful (Platt, Cristianini, & Shawe-Taylor, 2000). The DAG approach makes sense conceptually because it maps directly to the NAEP scale; that is, it distinguishes disfluent reading (Levels 1 and 2 in the NAEP scale) from fluent reading

(Levels 3 and 4 in the NAEP scale) and then makes finer distinctions (1 vs. 2 and 3 vs. 4). To implement the DAG strategy, we trained three classifiers. The first classifier was trained on samples from all classes and separated samples from Classes 1 and 2 and 3 and 4. This classifier was placed at the root of the tree, whereas two other classifiers, trained on samples from Classes 1 and 2 and 3 and 4, respectively, were placed on the leaves to make the finer-grained decisions. A detailed description of the classification scheme can be found in Bolaños et al. (2013).

Speech Recognition System

A total of 106 hr of read speech from three different children's speech corpora were used to train the recognition system. The recognizer was not trained on the corpus of read speech, described above, that was used to evaluate FLORA. We note that the system is *text independent*; that is, for new text passages, the system automatically generates the expected pronunciation(s) of each word in a text passage from a pronunciation dictionary.

The speech recognition system combines two main sources of information to produce a score for each word. These sources are (a) the score produced by matching the system's acoustic models for the expected sequence of phonemes in a word (based on a pronunciation dictionary) to the student's pronunciation of the word and (b) the probability of the word occurring in the text (the statistical language model, based on the co-occurrence of words in the text passage). These two sources of information are combined to produce the most likely hypothesis string given the speech input. Additionally, phone-level alignments from each of the 1-min recordings were generated automatically for feature extraction purposes. Two complementary speaker adaptation techniques were used in order to tailor the speaker-independent acoustic models to the speech characteristics and vocal tract length of each speaker.

Comparison Between Automated and Human Assessments of ORF

Human Scoring of Recorded Sessions

In order to evaluate the ability of FLORA to produce reliable WCPM scores, each of the 783 one-min recordings in the evaluation corpus was scored independently by two former elementary school teachers. Each teacher had more than a decade of experience administering reading assessments to elementary school children. The scorers were able to listen to, review, and modify their judgments within each recording until they were satisfied with their WCPM score. Thus, they were allowed to listen to the recording more than once.

Additionally, each of the 783 recordings was scored from 1 to 4 using the NAEP ORF scale by at least two independent scorers, who were former elementary school teachers with experience assessing reading proficiency. A set of 70 stories of the total 783 stories were scored by the five available teachers, whereas the other recordings were scored by just two of them, which were randomly assigned to each scorer. A training session was scheduled before the scoring process to review the NAEP scoring instructions and unify criteria. The judges first listened to passages rated by two experienced researchers whose area of expertise is expressive reading (Paula Schwanenflugel and Melanie Kuhn).

The teachers who scored the stories then rated these passages and compared their ratings with the experts. The teachers then rated several additional passages and reviewed their ratings based on the definitions of each of the NAEP levels. The training was concluded when the teachers' level of agreement approximated the agreement exhibited by the two experts.

For the actual scoring of the evaluation corpus, the judges listened to each 60-s story in 20-s intervals and provided a 1–4 rating for each interval. The NAEP ORF scale (Daane et al., 2005) comprises four levels from less to more fluent. Level 1 is characterized by word-by-word reading, Level 2 by reading using two-word phrases with some three- or four-word groupings, and Level 3 is characterized by a majority of three- or four-word phrase groups while preserving the syntax of the author. Readers at Level 4 produce larger, meaningful phrase groups with expressive interpretation. Finally, scorers attached a global NAEP score to the recording based on the NAEP scores assigned to each 20-s segment. The global score was based on a review of the scores and their best judgment rather than using a deterministic method like the mean or mode.

Assessment of Reading Accuracy and Automaticity

Table 1 shows the means and standard deviations (between parentheses) for accuracy, words per minute (WPM), and WCPM scores for the human scorers and FLORA. Statistics are shown per reading level for students in the four schools. As noted above, although the evaluation data were collected from students from Grades 1 to 4, about 20% of the time, teachers requested that specific students be presented with text passages either one or two levels below or above the student's grade level, resulting in reading levels for text passages from Grades 1 to 6. In Table 1, accuracy is expressed in percentages and WPM, which measures fluency from the perspective of speed-ignoring accuracy, and the score is based on the average across the two human scorers for each recording. It can be seen that accuracy (percentage of words read correctly) is higher for higher grade levels, from 70.3% for first grade to 92.6% and 90.5% for fifth- and sixth-grade levels, respectively. WPM are displayed in Column 5 for each grade level; as expected, they are highly correlated, with WCPM measured by human scorers (Column 5); however, WCPM computed by FLORA (Column 7) are much closer to human WCPM scores (Column 6) than WPM.

A major result can be observed by comparing the WCPM scores from the human scorers and FLORA, which present very similar

distributions (means and standard deviations). In addition, we observed very similar distributions of WCPM scores from humans and FLORA within each of the nine classrooms in which we conducted the study, even for classrooms in schools in which the majority of students spoken Spanish as their first language and were officially designated as English learners.

Column 8 shows the expected number of WCPM for each grade level according to Hasbrouck and Tindal (2006) reading norms. It can be seen in the table that students were assigned by teachers to reading levels at which they read around the 50th percentile. We believe that there is no credible evidence to link higher WCPM scores to improved comprehension, but there is substantial support for the need for readers to have an accuracy and rate (WCPM score) in the range of the 50th percentile to support both comprehension and motivation.

Another pattern of results is revealed by examining the numbers in Column 9, which shows the mean difference in WCPM scores for the two human scores for the recordings in each classroom, and the numbers in Column 10, which shows the mean difference between the averaged human scores and FLORA for each classroom. Note that differences in WCPM scores are expressed in absolute value. Viewing the numbers in Column 9 reveals the remarkable agreement between the two human scorers (1.2-WCPM difference across all schools) and the low variance. Across all recordings, the mean difference between FLORA and the averaged human scores was 3.6 words, whereas the mean difference between human scores was 1.2 words.

Figure 1a displays a scatter plot of the WCPM scores from the two human scorers for all recordings, whereas Figure 1b displays a scatterplot of the WCPM scores from FLORA with respect to the average human scores for all recordings. If agreement were perfect, all points would lie on the diagonal. These figures show the strong agreement between WCPM scores for human scorers on each recording, and the very good agreement between FLORA and the human scores, with relatively few outliers.

We were interested in determining whether FLORA might be a useful tool for providing WCPM scores that could be used as one valuable indicator, along with other measures, to identify students who are at risk for failing to learn to read. One way to do this is to compare human and FLORA WCPM scores with the national reading norms developed by Hasbrouck and Tindal (2006), which measured WCPM scores, for first- through sixth-grade students during each trimester of a school year. The interrater agreement in the task of mapping recorded stories to percentiles was 0.97 for the

Table 1

Summary of Accuracy, WPM, and WCPM According to Human Scorers (H) and FLORA (F). Expected WCPM (E) Are Also Shown

Level	Stu.	Rec.	Acc. (%)	H-WPM	H-WCPM	F-WCPM	E-WCPM	H-diff	FH-diff
1	68	171	70.3 (19.7)	54.6 (25.5)	41.9 (26.4)	42.5 (25.9)	53	1.2 (1.8)	2.7 (2.7)
2	97	242	84.6 (10.1)	99.3 (31.9)	85.7 (33.1)	86.1 (31.8)	89	1.2 (2.0)	3.8 (4.4)
3	52	128	87.3 (7.6)	113.4 (28.1)	100.1 (29.6)	101.6 (28.0)	107	1.2 (1.4)	3.6 (2.8)
4	59	147	87.4 (8.1)	124.4 (26.6)	109.9 (27.3)	112.7 (27.3)	123	1.3 (1.8)	4.1 (3.1)
5	30	76	92.6 (3.6)	156.9 (26.4)	145.6 (26.6)	145.6 (24.5)	139	1.1 (1.2)	4.6 (4.5)
6	7	19	90.5 (14.1)	145.9 (46.1)	137.3 (49.6)	137.4 (49.1)	150	1.5 (2.0)	2.8 (2.6)
All	313	783	83.3 (14.0)	103.3 (42.2)	90.1 (43.1)	91.1 (42.6)		1.2 (1.8)	3.6 (3.6)

Note. WPM = words per minute; WCPM = words correct per minute; FLORA = FLuent Oral Reading Assessment; Stu. = number of students; Rec. = number of recordings; Acc. = accuracy; H-diff = difference between the human scorers; FH-diff = difference between the FLORA and human scorers.

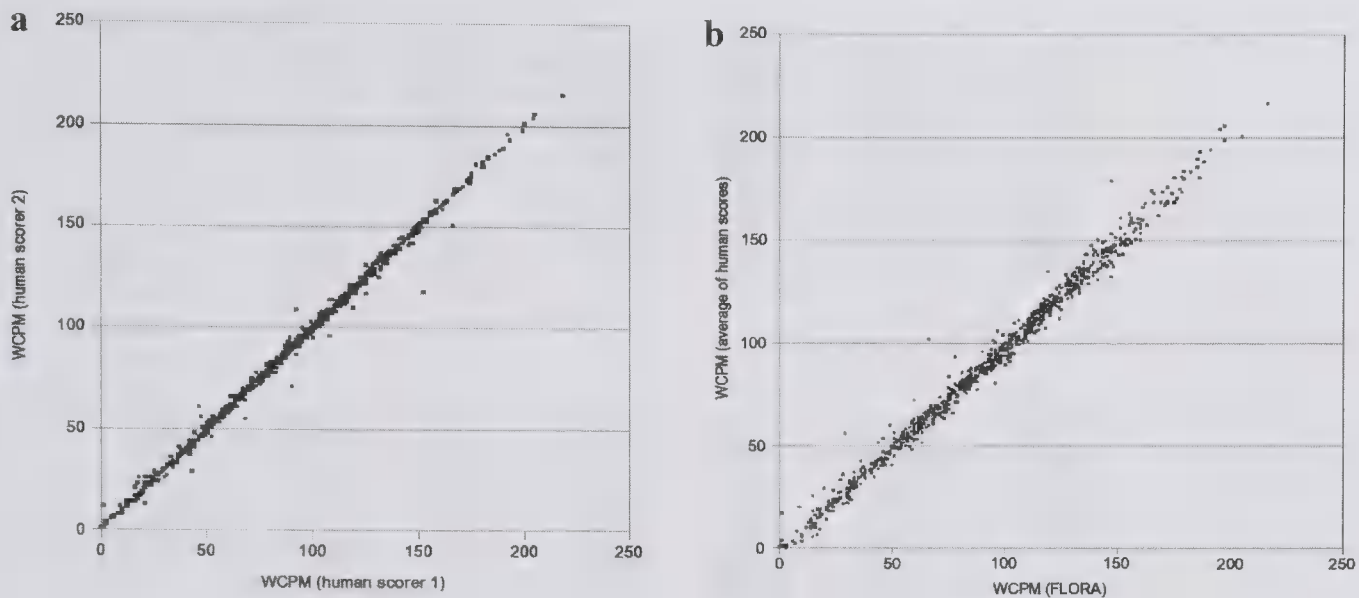


Figure 1. Correlation between WCPM scores produced by two independent human scorers (a) and between FLORA and the average of the two independent human scorers (b) for each of the 1-min recordings assessed. WCPM = words correct per minute; FLORA = FLuent Oral Reading Assessment.

human scorers and 0.89 between FLORA and each of the human scorers. The interrater agreement in the task of mapping recorded stories above and below the 50th percentile (which is used normally as a reference to identify at-risk students) was 0.98 for the human scorers and 0.92 between FLORA and each of the human scorers. Agreement was computed using the weighted kappa coefficient (κ) (Cohen, 1968), which is suitable for ordinal categories. In sum, the interhuman agreement and the FLORA to human agreement is very close, which means that FLORA performs well at identifying students who might require additional reading assessments and instruction.

Assessment of Expressive Reading

In this section, we present results on assessing expressive oral reading using FLORA. First, we briefly analyze the classification accuracy for the lexical and prosodic features proposed in relation to human assessments. We then analyze agreement and correlation between human scores and FLORA's automatic scoring system using the NAEP scale.

Classification accuracy. In order to derive the most effective combination of features to assess expressive reading, we measured the classification accuracy (percentage of recordings that FLORA assigned the same label than the human labelers) of FLORA on the corpus described above. Each recording was labeled by FLORA according to the NAEP scale, and labels were compared with those from all the available human labelers. We note that there exists an upper bound to the classification accuracy that can be attained by the classifier. The reason is that whenever the human raters score the same recording differently, there is an unrecoverable classification error.

Results showed that both lexical and prosodic features contributed similarly to the classification accuracy for the NAEP-2 (disfluent vs. fluent) task (89.27% and 89.02%, respectively). This can be initially considered an unexpected result because lexical aspects like the number of words read correctly are expected to dominate the discrimination between fluent and nonfluent readers. However,

it is important to note that some of the prosodic features defined in this study are highly correlated with the lexical features. For example, it is obvious that the number of words correctly read in a 1-min reading session should correlate highly with the average duration of a silence region or the number of filled pauses made.

For both the NAEP-2 and NAEP-4 tasks, lexical and prosodic features provided complementary information that led to improved classification accuracy when combined. For the NAEP-4 tasks, lexical features seem to have a dominant role (73.24% and 69.73%, respectively). We attribute this to the WCPM score, which is taken as a lexical feature; this score by itself provides a 71.78% accuracy for the NAEP-4 task. As expected, the automatically computed WCPM, which comprises two of the three reading fluency cornerstones (accuracy and rate), plays a fundamental role. In particular, the combination resulted in accuracies of 90.72% and 75.87% for the NAEP-2 and NAEP-4 tasks, respectively. Finally, note that the distribution of recordings across the NAEP levels according to humans and machine was very similar.

Interrater Agreement and Correlation

In this section, we present interrater agreement and correlation results for the best system from the previous section (multilabel training using all the features). Table 2 shows the interrater agreement for the tasks of classifying recordings into the broad NAEP categories (fluent vs. nonfluent; NAEP-2), or the four levels of expressiveness using the NAEP rubric (NAEP-4). For the NAEP-2 task, the interrater agreement was measured using Cohen's kappa coefficient (κ) (Cohen, 1960); $p(a)$ is the probability of observed agreement, whereas $p(e)$ is the probability of chance agreement.

For the NAEP-4 task, we measured the interrater agreement using the weighted kappa coefficient (κ) (Cohen, 1968), which is more suitable for ordinal categories given that it weights disagreements differently depending on the distance between the categories (we used linear weightings). As a complementary metric for this task, we computed the Spearman's rank correlation coefficient (Spearman, 1904). In a number of classification problems, like

Table 2
Interrater Agreement and Correlation Coefficients on the NAEP Scale

Scorer	# recordings	NAEP-2			NAEP-4	
		p(a)	p(e)	κ	κ	ρ
Human 1	571	0.87	0.50	0.73	0.66	0.80
Human 2	391	0.90	0.50	0.80	0.69	0.81
Human 3	698	0.87	0.50	0.74	0.68	0.81
Human 4	799	0.86	0.50	0.71	0.69	0.81
Human 5	367	0.86	0.50	0.71	0.68	0.80
FLORA	1,776	0.94	0.50	0.84	0.77	0.86

Note. NAEP = National Assessment of Educational Progress; FLORA = FLuent Oral Reading Assessment.

emotion classification, the data are annotated by a group of human raters who may exhibit consistent disagreements on similar classes or similar attributes. In such classification tasks, it is inappropriate to assume that there is only one correct label because different individuals may consistently provide different annotations (Steidl, Levit, Batliner, Nöth, & Niemann, 2005). Although the NAEP scale is based on clear descriptions of reading behaviors at each of four levels, children's reading behaviors can vary across these descriptions while reading, and individuals scoring the stories may differ consistently in how they interpret and weight children's oral reading behaviors. For this reason, we believe that examining correlations between human raters and between human raters and the machine classifiers is a meaningful and useful metric for this task.

Each row in Table 2 shows the agreement and correlation coefficients of each rater with respect to the other raters (excluding FLORA in the case of the human raters; note that not all the scorers scored the same number of recordings). In order to interpret the computed kappa values, we have used as a reference the interpretation of the kappa coefficient provided in Landis and Koch (1977), which attributes *good* agreement to kappa values within the interval (0.61–0.80) and *very good* agreement to higher kappa values (0.81–1.00). According to this interpretation, Table 2 reveals that (a) there is *good* interhuman agreement for both the NAEP-2 and NAEP-4 tasks, (b) there is *good* FLORA-to-human agreement for the NAEP-4 task, and (c) there is *very good* FLORA-to-human agreement for the NAEP-2 task. It can be observed that the kappa agreement between FLORA and the humans is higher than the agreement between each human scorer and the rest of the human scorers. This is true for both the NAEP-2 and NAEP-4 tasks. This difference in agreement is statistically significant, which indicates the ability of the proposed features and

classification scheme to provide a useful method to automatically assess expressive oral reading using the NAEP scale.

In terms of the Spearman's rank correlation coefficient (ρ), we obtained relatively strong interhuman correlation (.80–.81) and an even stronger machine-to-human correlation (.86) in the NAEP-4 task. This indicates that NAEP scores from every pair of scorers are closely related, which is consistent with the weighted kappa values obtained.

In Table 3, we display cross-tabs of agreement and disagreement between humans and between FLORA and humans (in percentages). In both cases, most of the data lie in the main diagonal, and we believe that there are no obvious biases between humans and FLORA.

Connection Between Reading Accuracy, Reading Rate, and Expressive Reading

We conducted a set of analyses to gain insights into the relationship between the two main measures of ORF, WCPM, and expressiveness. These analyses are displayed in Figure 2a and 2b. In each panel of the figure, we sorted students according to their WCPM percentile using the Hasbrouck and Tindal (2006) norms. Thus, the leftmost bar of each panel represents students with WCPM scores below the 10th percentile, whereas the rightmost bar shows students in the 90th percentile. Figure 2a displays percentile assignments based on average human scorers rating, and Figure 2b displays percentile assignments based on FLORA WCPM estimates. The tones of gray within each bar indicate the percentage of students at each NAEP score; in Figure 2a, these numbers are based on the NAEP scores assigned by the human scorers, and in Figure 2b these numbers were assigned by FLORA.

Table 3
Cross-Tabs of Agreement/Disagreement Between FLORA and Human-Generated NAEP Scores (in %)

		FLORA						Human			
		1	2	3	4			1	2	3	4
Human	1	16.6	2.9	0.1	0	Human	1	14.9	4.2	0.1	0
	2	3.5	21.3	3.9	0.2		2	4.4	19.6	5.7	0.2
	3	0	3.9	32.4	5.6		3	0.1	7.2	27.7	5.2
	4	0	0	3	6.6		4	0	0	4.7	6.1

Note. FLORA = FLuent Oral Reading Assessment; NAEP = National Assessment of Educational Progress.

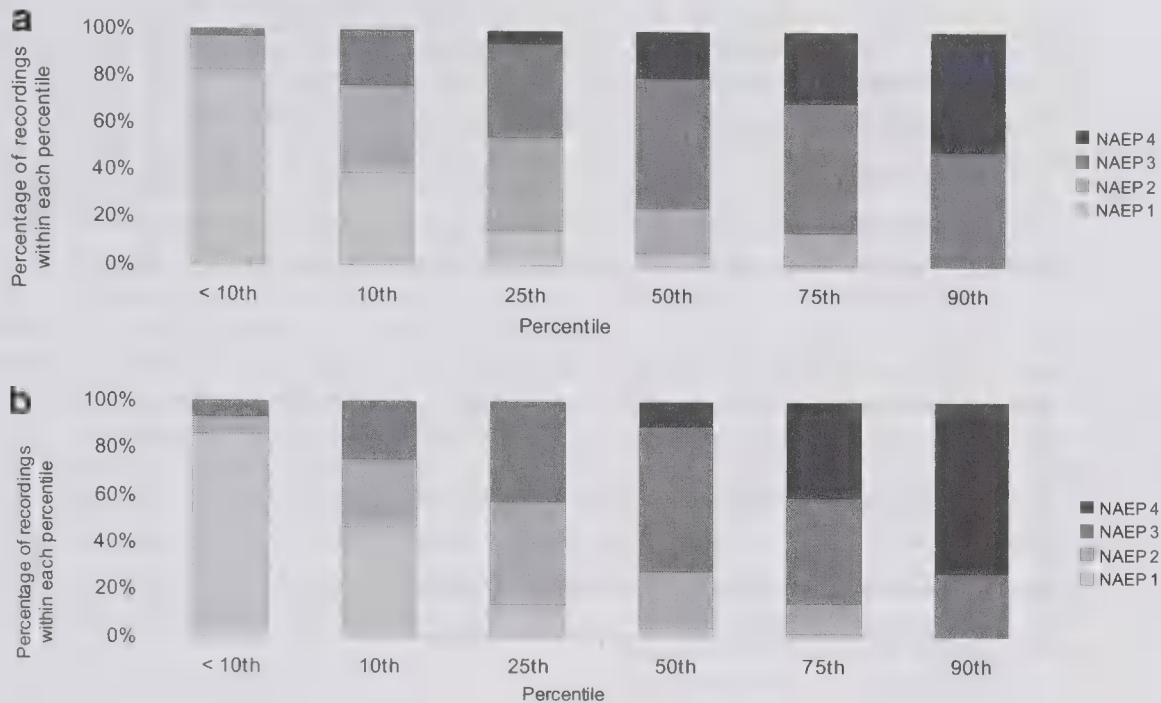


Figure 2. a: Distribution of recordings across the NAEP scale for each WCPM percentile according to human scorers. b: Distribution of recordings across the NAEP scale for each WCPM percentile according to FLORA. NAEP = National Assessment of Educational Progress; WCPM = words correct per minute.

It is clear from this figure that recordings in the highest percentiles (highest reading accuracy and rate) correspond to more expressive readers (higher levels in the NAEP scale). For example, all of the recordings for students in the 90th percentile based on WCPM were assigned to Levels 3 and 4 in the NAEP scale. Moreover, about 97.0% of the recordings below the 10th percentile were assigned to Levels 1 and 2 in the NAEP scale. Figures 2a and 2b reveal several interesting patterns: A significant percentage of recordings placed below the 50th percentile (which might be used to identify students in need for fluency support) were placed in the higher levels of the NAEP scale according to our expert human annotators (3.08%, 24.02%, and 45.19% for recordings below the 10th percentile, in the 10th percentile, and in the 25th percentile, respectively). This means that there are a number of speakers who, despite reading below the expected rate according to the percentiles published by Hasbrouck and Tindal (2006), read with appropriate/good expression and would be considered fluent readers according to the NAEP scale. Another interesting observation is that a significant percentage of recordings placed above the 50th percentile were assigned to the lower levels in the NAEP scale by our expert human annotators. Those recordings likely correspond to speakers who are reading for speed rather than for comprehension in order to get as many words read as possible within the 1-min session. In particular, 24.88% of the recordings in the 50th percentile were assigned to Levels 1 and 2 in the NAEP scale (nonfluent), whereas 13.92% of the recordings in the 75th percentile were assigned to those levels. We note that the instructions provided to students before recording stories emphasized the importance of reading the text naturally, rather than as fast as they could; these percentage might have been higher if we had not emphasized reading naturally in the instructions. These observations suggest that measuring both expressiveness and WCPM is likely to be both informative and beneficial to understanding

individual student's oral reading abilities. Finally, we note that Figure 2b, which is analogous to Figure 2a but was built using FLORA scores, presents very similar information.

Discussion and Conclusions

We investigated the automatic assessment of ORF in children's speech according to two standard rubrics: WCPM (to measure accuracy and rate) and the NAEP Expressiveness scale. Compared with human scoring of WCPM and expressiveness on 783 one-min recordings of children reading grade-level text passages, results show that automatically generated WCPM scores differ by an average of 3.5 words with respect to the human-average score for each recorded story, whereas humans differ by an average of 1.5 words for each story.

For expressiveness, FLORA had an accuracy of 90.93% classifying recordings according to the binary NAEP scale ("fluent" vs. "nonfluent") and 76.05% on the more difficult 4-point NAEP scale. According to the classification of kappa strength proposed by Landis and Koch (1977), the kappa agreement for both NAEP-2 and NAEP-4 tasks between each human scorer and the rest of the human scorers was *good*, whereas the kappa agreement between the machine and the human scorers was *good* and *very good*, respectively. In addition, the kappa agreement between FLORA and each human scorer was always significantly higher than the kappa agreement between the human scorers. In terms of the Spearman's rank correlation coefficient (ρ), correlation between the machine and each human scorer was always significantly higher than the correlation between human scorers.

The results of the research reveal that speech recognition and machine learning systems can produce accurate assessments of WCPM and expressiveness that approach (WCPM) or exceed human performance. Without question, the results of the WCPM

scores reported above can be improved substantially in the near future using known ASR solutions, such as collecting more training data to model children's speech patterns. For example, Ver-gyri, Lamel, and Gauvain (2010) reported that accent-dependent acoustic modeling (which implies training/adapting on data from the target accent) produces a significant increase in recognition performance compared with accent-independent modeling. In a recent study that we conducted on 191 native Spanish children learning to read English text in Spanish schools (Bolaños, Elhazaz, Ward, & Cole, 2012), we determined experimentally that statistical models trained on speech from the target population were significantly more accurate than models trained on native English children. Results from that study showed a mean difference in WCPM scores of 5.49 and 4.96, respectively, between FLORA and each of the human scorers, whereas the mean difference between the human scorers was about 5.92 words.

Perhaps the major limitation of this study is the relatively small number of students (313) used in our research. To fully demonstrate the feasibility and validity of a fully automatic assessment of ORF, speech data during oral reading of leveled texts must be collected for a large and diverse population of students at different grade levels, representing students with different dialects and accents. The system must also be tested with data collected from many different classrooms or computer labs to model the acoustic environments and the realities of real-world use.

Toward Valid Automatic Assessment of ORF

We believe there are great potential benefits of incorporating measures of expressiveness into assessments of ORF. One of the major criticisms of using WCPM to measure individual student's improvements in reading over time (i.e., in response to instruction) is that students strive to read texts as quickly as possible in order to increase their WCPM scores, which teachers often set as learning targets within a reading instruction program. When a student's ability is measured in terms of how quickly he or she can read the words in a text, teachers and students learn to focus on reading fast, rather than reading the text at a normal reading rate with intonation and phrasing that communicates the meaning of text, and thus reflects its comprehension by the student. Fast readers have shorter segment durations, muted stress marking, and reduced phrase-final bracketing than slow readers, so the normal comprehension benefits children might experience by reading with good prosody may not be derived by students who are trying to read fast (Benjamin & Schwanenflugel, 2010; Kuhn et al., 2010). In sum, the emphasis on speed that can result from using WCPM as the only measure of ORF may undermine the goal of helping students develop strategies for reading with deep understanding.

Incorporating measures of expressiveness into assessments of ORF could mitigate this problem. One can easily imagine a weighted measure of ORF that combines WCPM and expressiveness estimates, such that students receive the highest score when the words in a text are read at a natural speaking rate with prosody appropriate to the discourse structure of the text. In fact, some rating systems of reading expressiveness such as the Multidimensional Fluency Guide (Rasinski, Rikli, & Johnston, 2009) already do this.

One of the major benefits of the automated scoring of reading prosody by FLORA that neither the NAEP nor the other various

teacher rating systems for evaluating reading fluency have is that these reading fluency scales have not (as yet) been grounded in research on reading prosody. We do not know whether the ratings obtained using these scales would be spectrographically valid, that is, that children rated as expressive on these scales would be the same ones who would appear expressive when their readings are viewed on a spectrogram. Because the features used in FLORA to classify expressive reading were derived directly from spectrographic measures derived from children's speech (Kuhn et al., 2010), FLORA can make this claim. Conversely, because the teacher NAEP ratings match the spectrographic distinctions made by FLORA, FLORA has also served to validate teacher impressions of reading prosody as determined by the NAEP. In sum, fully automatic assessment or ORF that combines its three components appears to be feasible with today's technologies. Additional research is needed to determine how to use these measures to provide the most useful feedback to teachers and students to assess students' reading abilities and inform instruction.

References

- Benjamin, R., & Schwanenflugel, P. J. (2010). Text complexity and oral reading prosody in young readers. *Reading Research Quarterly*, 45, 388–404.
- Black, M., Tepperman, J., Lee, S., & Narayanan, S. (2008). *Estimation of children's reading ability by fusion of automatic pronunciation verification and fluency detection*. Proceedings of Interspeech, Brisbane, Australia.
- Bolaños, D. (2012, December). *The Baviaca open-source speech recognition toolkit*. In Proceedings of IEEE Workshop on Spoken Language Technology (SLT), Miami, FL.
- Bolaños, D., Cole, R. A., Ward, W., Borts, E., & Svirsky, E. (2011). FLORA: Fluent oral reading assessment of children's speech. *ACM Transactions on Speech and Language Processing*, 7, 1–19. doi:10.1145/1998384.1998390
- Bolaños, D., Cole, R. A., Ward, W., Tindal, G., Schwanenflugel, P., & Kuhn, M. (2013). Automatic assessment of expressive oral reading. *Speech Communication*, 55, 221–236. doi:10.1016/j.specom.2012.08.002
- Bolaños, D., Elhazaz, P., Ward, W., & Cole, R. (2012). Automatic assessment of oral reading fluency for native Spanish ELL children. In *Proceedings of WOCCI 2012: Workshop on Child, Computer and Interaction: Satellite Event of INTERSPEECH*.
- Bolaños, D., Ward, W. H., Wise, B., & Vuuren, S. V. (2008, September). *Pronunciation error detection techniques for children's speech*. Paper presented at INTERSPEECH 2008, Brisbane, Australia.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46. doi:10.1177/001316446002000104
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213–220. doi:10.1037/h0026256
- Compton, D. L., & Carlisle, J. F. (1994). Speed of word recognition as a distinguishing characteristic of reading disabilities. *Educational Psychology Review*, 6, 115–140. doi:10.1007/BF02208970
- Curtis, M. E. (1980). Development of components of reading skill. *Journal of Educational Psychology*, 72, 656–669. doi:10.1037/0022-0663.72.5.656
- Daane, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., & Oranje, A. (2005). *Fourth-grade students reading aloud: NAEP 2002 special study of oral reading* (NCES 2006–469). Washington, DC: U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics.

- Duong, M., Mostow, J., & Sitaram, S. (2011). Two methods for assessing oral reading prosody. *ACM Transactions on Speech and Language Processing*, 7.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5, 239–256. doi:10.1207/S1532799XSSR0503_3
- Good, R. H., & Kaminski, R. A. (2002). *Dynamic indicators of basic early literacy skills*. Eugene, OR: Institute for the Development of Educational Achievement.
- Hagen, A., & Pellom, B. (2005, April). *A multi-layered lexical-tree based recognition of subword speech units*. Paper presented at the Second Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznan, Poland.
- Hagen, A., Pellom, B., & Cole, R. (2007). Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication*, 49, 861–873. doi:10.1016/j.specom.2007.05.004
- Hasbrouck, J., & Tindal, G. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher*, 59, 636–644. doi:10.1598/RT.59.7.3
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, 2, 127–160. doi:10.1007/BF00401799
- Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review*, 34, 9–26.
- Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C. L., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95, 719–729. doi:10.1037/0022-0663.95.4.719
- Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, 102, 652–667. doi:10.1037/a0019643
- Kuhn, M. R., Schwanenflugel, P. J., & Meisinger, E. B. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly*, 45, 230–251. doi:10.1598/RRQ.45.2.4
- Kuhn, M. R., & Stahl, S. (2000). *Fluency: A review of developmental and remedial practices*. Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.
- LaBerge, D., & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293–323. doi:10.1016/0010-0285(74)90015-2
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. doi:10.2307/2529310
- Lesgold, A. M., & Resnick, L. B. (1982). How reading disabilities develop: Perspectives from longitudinal study. In J. P. Das, R. Mulcahy, & A. Wall (Eds.), *Theory and research in learning disability*. New York, NY: Plenum Press.
- Marston, D. (1989). Curriculum-based measurement: What is it and why do it? In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York, NY: Guilford Press.
- Miller, J., & Schwanenflugel, P. J. (2008). A longitudinal study of the development of reading prosody as a dimension of oral reading fluency in early elementary school children. *Reading Research Quarterly*, 43, 336–354. doi:10.1598/RRQ.43.4.2
- Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., & Tobin, B. (2003). Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, 29, 61–117. doi:10.2190/06AX-QW99-EQ5G-RDCF
- National Reading Panel, National Institute of Child Health and Human Development. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development, National Institutes of Health.
- Perfetti, C. (1985). *Reading ability*. Oxford, England: Oxford University Press.
- Platt, J. C., Cristianini, N., & Shawe-Taylor, J. (2000). Large margin DAGs for multiclass classification. *Advances in Neural Information Processing Systems*, 12, 547–553.
- Rasinski, T., Rikli, A., & Johnston, S. (2009). Reading fluency: More than automaticity? More than a concern for the primary grades? *Literacy Research and Instruction*, 48, 350–361. doi:10.1080/19388070802468715
- Reeder, K., Shapiro, J., & Wakefield, J. (2007). The effectiveness of speech recognition technology in promoting reading proficiency and attitudes for Canadian immigrant children In *Proceedings of the 9th European Conference on Reading*. Berlin, Germany: IDEC.
- Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading comprehension. *School Psychology Quarterly*, 20, 304–317. doi:10.1521/scpq.2005.20.3.304
- Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology*, 46, 343–366. doi:10.1016/j.jsp.2007.06.006
- Samuels, J. (1985). *Automaticity and repeated reading*. Lexington, MA: Lexington Books.
- Scarborough, H. S. (1998). Early identification of children at risk for reading difficulties: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75–199). Timonium, MD: York Press.
- Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *Elementary School Journal*, 107, 429–448. doi:10.1086/518622
- Schwanenflugel, P. J., Meisinger, E. B., Wisenbaker, J. M., Kuhn, M. R., Strauss, G. P., & Morris, R. D. (2006). Becoming a fluent and automatic reader in the early elementary school years. *Reading Research Quarterly*, 41, 496–522.
- Shinn, M. (1998). *Advanced applications of curriculum based measurement*. New York, NY: Guilford Press.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101. doi:10.2307/1412159
- Steidl, S., Levit, M., Batliner, A., Nöth, E., & Niemann, H. (2005). “Of all things the measure is man”: Automatic classification of emotions and interlabeler consistency. In *Proceedings ICASSP* (Vol. 1, pp. 317–320). doi:10.1109/ICASSP.2005.1415114
- Vapnik, V. (1995). *The nature of statistical learning theory*. New York, NY: John Wiley & Sons.
- Vergyri, D., Lamel, L., & Gauvain, J. L. (2010, September). *Automatic speech recognition of multiple accented English data*. Paper presented at INTERSPEECH 2010, Makuhari, Japan.
- Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, 41, 85–120. doi:10.1177/00224669070410020401
- Wolf, M. (1999). What time may tell: Towards a new conceptualization of developmental dyslexia. *Annals of Dyslexia*, 49, 1–28. doi:10.1007/s11881-999-0017-x

Received December 16, 2011

Revision received November 19, 2012

Accepted December 10, 2012 ■

Cognitive Anatomy of Tutor Learning: Lessons Learned With SimStudent

Noboru Matsuda, Evelyn Yarzebinski,
Victoria Keiser, Rohan Raizada,
and William W. Cohen
Carnegie Mellon University

Gabriel J. Stylianides
University of Oxford

Kenneth R. Koedinger
Carnegie Mellon University

This article describes an advanced learning technology used to investigate hypotheses about learning by teaching. The proposed technology is an instance of a teachable agent, called *SimStudent*, that learns skills (e.g., for solving linear equations) from examples and from feedback on performance. *SimStudent* has been integrated into an online, gamelike environment in which students act as “tutors” and can interactively teach *SimStudent* by providing it with examples and feedback. We conducted 3 classroom “in vivo” studies to better understand how and when students learn (or fail to learn) by teaching. One of the strengths of interactive technologies is their ability to collect detailed process data on the nature and timing of student activities. The primary purpose of this article is to provide an in-depth analysis across 3 studies to understand the underlying cognitive and social factors that contribute to tutor learning by making connections between outcome and process data. The results show several key cognitive and social factors that are correlated with tutor learning. The accuracy of students’ responses (i.e., feedback and hints), the quality of students’ explanations during tutoring, and the appropriateness of tutoring strategy (i.e., problem selection) all positively affected *SimStudent*’s learning, which further positively affected students’ learning. The results suggest that implementing adaptive help for students on how to tutor and solve problems is a crucial component for successful learning by teaching.

Keywords: learning by teaching, machine learning, *SimStudent*, teachable agent, tutor learning

It has been widely observed that students learn by teaching others (e.g., E. G. Cohen, 1994). Such an effect of learning by teaching (also known as the *tutor-learning effect*) has been empirically confirmed in many different domains for many different structures of peer tutoring with different ages and achievement levels (Roscoe & Chi, 2007). Despite a long-standing history of empirical studies on the tutor-learning effect, not enough is known about the cognitive and social theory of when, how, and why tutors learn (or fail to learn) by teaching.

A primary challenge in theory development for tutor learning is a lack of the *process data*, that is, a detailed record of interactions between tutors and tutees. Collecting rich process data from peer tutoring sessions can enable descriptions of tutoring activities at a fine level of granularity, such as dialogue between the tutor and the tutee, response accuracy, and timing and sequencing of actions. When combined with outcome data (e.g., test scores), this detailed information can allow further exploration of elements of cognitive and social theory of tutor learning. However, such process data are rarely available. Roscoe and Chi (2007) reported that only six out of thousands of related articles report both outcome and process data. An obvious reason for the lack of process data is the difficulty in collecting such data during a study in which human students tutor their peers.

In their meta-analysis of prior research, Roscoe and Chi (2007) summarized potential flaws in program design and implementation that might have impacted the tutor-learning effect. One way to avoid such flaws is to better understand the process of tutor learning and to provide appropriate facilities for the tutors. Knowledge gained from combined process and outcome data can aid iterative design engineering of more effective learning by tutoring.

To help advance the cognitive and social theory of tutor learning, we have developed a synthetic pedagogical agent as a tutee that students can interactively tutor. Such a pedagogical agent is often called a *teachable agent*, which in our case is named *SimStudent* (Matsuda, Cohen, Sewall, Lacerda, & Koedinger, 2007). *SimStudent* engages in genuine machine learning to learn proce-

This article was published Online First September 9, 2013.

Noboru Matsuda, Evelyn Yarzebinski, Victoria Keiser, Rohan Raizada, and William W. Cohen, School of Computer Science, Carnegie Mellon University; Gabriel J. Stylianides, Department of Education, University of Oxford, Oxford, England; Kenneth R. Koedinger, School of Computer Science, Carnegie Mellon University.

The research reported here was supported by National Science Foundation Award No. DRL-0910176 and the Institute of Education Sciences, U.S. Department of Education Grant R305A090519 awarded to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. This work is also supported in part by the Pittsburgh Science of Learning Center, which is funded by National Science Foundation Award No. SBE-0836012.

Correspondence concerning this article should be addressed to Noboru Matsuda, School of Computer Science, Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: Noboru.Matsuda@cs.cmu.edu

dural problem-solving skills. SimStudent has been integrated into an online, gamelike environment called *APLUS* (Artificial Peer Learning environment Using SimStudent).

With *APLUS* and SimStudent, we have conducted three tightly controlled in vivo studies for middle-school students learning algebra linear equations (Matsuda, Cohen, et al., 2012; Matsuda et al., 2011; Matsuda, Yarzebinski, et al., 2012). Solving linear equations is a critical area in the early algebra curriculum, yet many secondary school students experience great difficulty making the transition from arithmetic to algebra, especially in learning how to solve equations (see, e.g., Bednarz & Janvier, 1996; Filloy & Rojano, 1989; Kieran, 1992; Linchevski & Herscovics, 1996). Developing an effective intervention to learn equation solving thus has an urgent, practical need as well.

In this article, we investigated the following research questions:

Question 1: Does SimStudent actually learn how to solve equations when tutored by students in an authentic classroom setting? Accordingly, do students learn by teaching SimStudent?

Question 2: How do tutor and tutee learning correlate with each other?

Question 3: When and how do students learn or fail to learn by teaching SimStudent?

To answer these questions, we conducted in-depth analyses across three in vivo studies to understand the underlying cognitive and social factors that contribute to tutor learning. These analyses benefited from both outcome and process data.

In the rest of the article, we first provide a survey of prior research on the tutor-learning effect and the teachable agent technology. We then introduce SimStudent and *APLUS*, with a technical overview of how SimStudent acts as a teachable agent. Next, we explain how students interactively tutor SimStudent and provide an overview of the data analysis, which includes empirical data collected from the three in vivo studies. We then discuss how and when students learn or fail to learn by teaching SimStudent based on the process and outcome data. We conclude with a discussion of directions for future research based on the lessons learned from our studies.

The Tutor-Learning Effect

The tutor-learning effect has been studied for many years (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; P. A. Cohen, Kulik, & Kulik, 1982; Devin-Sheehan, Feldman, & Allen, 1976; Gartner, Kohler, & Riessman, 1971; Graesser, Person, & Magliano, 1995) and for different age groups, varying from elementary (Sharpley, Irvine, & Sharpley, 1983) to middle school (Jacobson et al., 2001; King, Staffieri, & Adelgais, 1998) to college (Annis, 1983; Topping, 1996). It has also been observed in various subject domains, including mathematics, reading, science, and social studies (P. A. Cohen et al., 1982; Cook, Scruggs, Mastropieri, & Casto, 1986; Mastropieri, Spencer, Scruggs, & Talbott, 2000; Mathes & Fuchs, 1994; Rohrbeck, Ginsburg-Block, Fantuzzo, & Miller, 2003), and in different forms of tutoring, including reciprocal tutoring (Palincsar & Brown, 1984), collaborative passage learning (Bargh & Schul, 1980), and small-group learning as opposed to peer-to-peer learning (Webb & Mastergeorge, 2003). It has also been demonstrated that tutors can learn by just preparing for teaching (Biswas et al., 2001).

Learning by teaching has been shown to be effective for minority populations. Robinson, Schofield, and Steers-Wentzell (2005) found that African American student tutors learned more from math peer tutoring than White students. Rohrbeck et al. (2003) found a larger effect size in groups with more than 50% minority enrollment than groups with lower minority enrollment. Other researchers found positive outcomes for students from underprivileged backgrounds (Greenwood, Delquadri, & Hall, 1989; Jacobson et al., 2001) and students with learning disabilities (Cook et al., 1986; Mastropieri et al., 2000).

Despite the fact that many experimental studies support the tutor-learning effect, the actual effect size has been known to be rather moderate (P. A. Cohen et al., 1982; Cook et al., 1986; Mastropieri et al., 2000; Mathes & Fuchs, 1994; Rohrbeck et al., 2003). The tutor-learning effect has been shown to be relatively more effective in math than reading. For example, P. A. Cohen et al. (1982) showed an effect size of .62 for math and .21 for reading, and Cook et al. (1986) showed an effect size of .67 for math and .30 for reading.

In sum, learning by teaching has the potential to be a successful intervention for a wide variety of student populations across many disciplines. It also has the potential to minimize the achievement gap between student demographic diversities. Despite the popularity of the tutor-learning effect, we lack an adequate cognitive theory of tutor learning. Understanding the underlying cognitive principles of tutor learning could facilitate the development of effective learning technologies and may improve on the rather small effect size of tutor learning.

Teachable Agent

There are a number of advantages of using a teachable agent technology to study the tutor-learning effect (e.g., VanLehn, Ohlsson, & Nason, 1994). First, it enables implementation of tight, precisely determined control conditions. For example, the variance of tutees can be controlled by having students teach the same version of the teachable agent. The teachable agent technology also allows researchers to control the competency of the tutee to see how it may affect tutor learning. Second, the teachable agent allows researchers to conduct peer-tutoring studies without the risk of harming tutees. Although nonexpert tutors have a greater chance of teaching inaccurate knowledge, previous studies showed that tutors often learned at the cost of tutee errors. Walker, Rummel, and Koedinger (2009) found that the amount of tutee errors had a significant positive correlation with tutor learning, whereas it had a significant negative correlation with tutee learning. Third, the teachable agent technology facilitates the collection of detailed process data showing interactions between the student and the agent, which is a major contribution of the current article.

There have been three major techniques used to build teachable agents: (a) Some teachable agents (TAs) solve problems using the shared knowledge that students create. Students using such *knowledge-sharing TAs* are often told that they teach the agent by directly providing the shared knowledge to the agent. For example, students teach Betty's Brain by drawing a concept map representing causal relationships between factors related to river ecology (Biswas, Leelawong, Schwartz, Vye, & The Teachable Agents Group at Vanderbilt, 2005; Leelawong & Biswas, 2008). (b) Another type of TA applies to the knowledge-tracing technique

that Cognitive Tutors use to diagnose students' competency (Ritter, Anderson, Koedinger, & Corbett, 2007). Such *knowledge-tracing TAs* are equipped with a set of skills to be learned. Some of the skills are set to be inactive at the beginning to provide the agent with limited competency to solve problems. As the student tutors the agent, the model-tracer identifies the skill that was tutored and activates the tutored skill so that the agent can apply it to future problems. Pareto, Arvemo, Dahl, Haake, and Gulz (2011) developed a knowledge-tracing TA for students to learn arithmetic concepts. (c) The last type of TA integrates machine-learning engines that allow the TA to learn skills dynamically, arguably more accurately reflecting the tutor-tutee interaction. As an example of such a *knowledge-learning TA*, Michie, Paterson, and Hayes (1989) developed the Math Concept Learning System with an inductive logic programming engine (called ID3) developed by Quinlan (1986) to induce rules from examples, which enabled it to learn math skills and solve equations. STEP (Simulated, Tutorable Physics Student) is another example of the knowledge-learning TA in Physics (Ur & VanLehn, 1995).

SimStudent is an example of a knowledge-learning TA, but has several distinctive characteristics compared with other TAs. First, SimStudent is one of a few TAs that have been intensively used in authentic classroom settings. Other such empirically well-validated agents include Betty's Brain (Biswas, Jeong, Kinnebrew, Sulcer, & Roscoe, 2010) and the TA developed by Pareto et al. (2011). Second, in contrast to other TAs, which have been largely implemented in declarative domains, SimStudent learns algebra content with a focus on procedural problem solving. Third, SimStudent is an instance of a TA with a *humanlike* learning capability (Li, Matsuda, Cohen, & Koedinger, 2011; Ohlsson, 2008). SimStudent performs inductive learning to interactively generalize examples provided by the student. Therefore, a naturalistic tutoring dialogue can occur between the student and SimStudent. Fourth, because SimStudent inductively learns skills from examples, it may learn skills incorrectly, depending on the prior knowledge it is given and the way the student tutors SimStudent. One such common source of incorrect learning stems from ambiguities in examples. To the best of our knowledge, SimStudent is the first TA that models students' incorrect *learning*. Because students generally learn both from correct and incorrect examples (Booth & Koedinger, 2008), observing a TA learning incorrectly may positively impact tutor learning.

Overview of the Data Analysis

To connect the outcome and process data to advance cognitive and social theories of the tutor-learning effect, data from three in vivo classroom studies have been analyzed to address the three research questions mentioned in the introduction. To measure SimStudent's learning, we used the process data showing how well SimStudent performed on the quiz. To measure students' learning (i.e., tutor learning), we used test scores as the outcome data. The correlation between SimStudents' and students' learning was analyzed using these two variables as well. We focused on a number of factors in the process data to analyze how and when SimStudents' and students' learning play out.

The Learning Environment: APLUS and SimStudent

Figure 1 shows an example screenshot of APLUS with SimStudent. The SimStudent avatar is visualized in the lower left corner. There have been three versions of SimStudent developed with different avatar images, as shown in Figure 2. Different versions of SimStudent have different functionalities to address different research questions as described later.

The initial version of SimStudent is called *Lucy* and is represented as a single static image (see Figure 2 i). The second version of SimStudent is called *Stacy* (see Figure 2 ii) and is capable of three facial expressions, including a thinking pose when SimStudent commits to learning, a happy expression when a problem is solved, and a neutral expression otherwise. The third version of SimStudent is called *Tomodachi* (see Figure 2 iii). Students can customize Tomodachi's avatar by changing the name, hairstyle, skin color, eyes, and shirt. Tomodachi is capable of the same three facial expressions as Stacy.

Overview of Tutoring Interaction

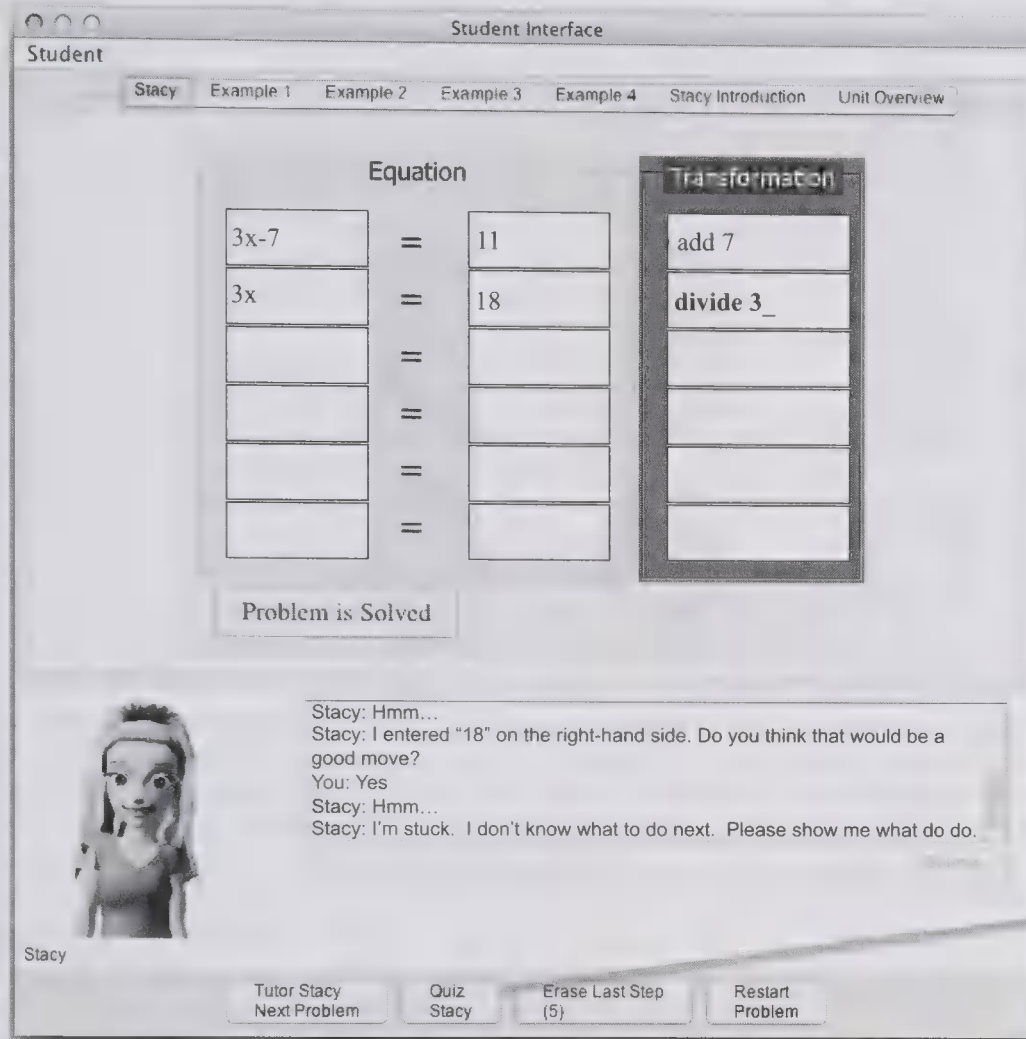
In APLUS, a student interactively tutors SimStudent with the following tutoring actions:

- Pose a problem for SimStudent to solve in the Tutoring Interface. In Figure 1a, the student entered " $3x - 7$ " and "11" in the first row of the equation table. SimStudent then attempts to solve the problem by applying learned productions and asking the student about the correctness of each step.
- Provide flagged (yes/no) feedback to SimStudent that shows the student's judgment on the correctness of SimStudent's steps. When the student provides negative feedback, SimStudent may make another attempt. In Figure 1a, SimStudent entered "18" in the second row, and asked whether the student thought it was a good move. The student then provided positive feedback.
- Provide help on what to do next. When SimStudent does not know what to do, SimStudent asks the student for help. To respond to the help request, the student demonstrates the next step in the tutoring interface. In Figure 1a, SimStudent got stuck after entering "18." In response, the student tutors SimStudent by showing it a possible next step, in this case entering "divide 3" for the transformation of the second row.
- Quiz SimStudent to gauge learning. Students may have SimStudent take (and retake) the quiz at any time during tutoring (see Figure 1c). Further details of the quiz are below.

There are also resources for students to review learning objectives in the unit overview and to review problem-solving procedures by studying worked-out examples. Clicking the different [Example] tabs displays complete examples in the Tutoring Interface. The [Unit Overview] tab provides a brief overview of the target unit (i.e., equations with variables on both sides), a model solution with elaborated explanations, and suggested problems for students to use when tutoring SimStudent.

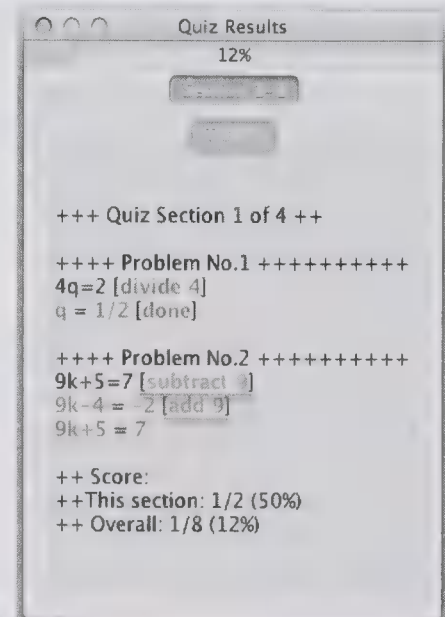
Quiz

In the classroom studies, students were told that their goal was to tutor SimStudent well enough so that SimStudent would pass a predefined quiz. The quiz has four sections each with two equation problems. There is a one-step equation (e.g., $3x = 6$), three



a) The APLUS interface

c) Summary of the quiz results



After SimStudent takes the quiz, a summary dialog window is shown. Incorrect steps are shown in red (but underlined in this figure).

b) SimStudent asking why a step she performed was incorrect

The student answering SimStudent's question by typing a free text in a chat box.

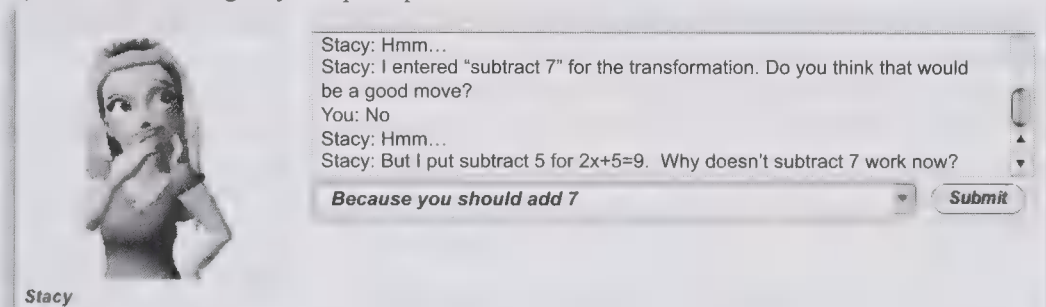


Figure 1. A screenshot of the Study II APLUS, the online gamelike learning environment in which the student can interactively tutor SimStudent. Students enter a problem on the first row of the Equation column. SimStudent attempts to solve the problem by entering steps (e.g., "3x" and "18" in this case). SimStudent asks the student about the correctness of the steps. When SimStudent cannot perform a step correctly, it asks the student for help. In this example, the student entered "divide 3" in the second row as a next step after SimStudent entered "18." SimStudent occasionally asks questions (b). In this example, SimStudent is asking for a reason why a step it performed is considered to be wrong. Students may have SimStudent take the quiz by clicking on the [Quiz Stacy] button. After SimStudent takes the quiz, a summary dialog window is shown (c). APLUS = Artificial Peer Learning environment Using SimStudent.

two-step equations (e.g., $-2x + 5 = 11$), and four equations with variables on both sides (e.g., $3 - 2x = 5x + 7$). SimStudent takes the quiz section by section, and must correctly solve both problems in each section to proceed to the next section.

After SimStudent takes the quiz, the overall results and correctness of the steps are displayed in a different window, as shown in Figure 1c. An embedded Cognitive Tutor Algebra program (Ritter et al., 2007) grades the quiz results. The Cognitive Tutor is invisible to students.

The quiz problems were randomly ordered for Study I, but they were ordered on the basis of increasing difficulty level for Studies II and III. The quiz problems were fixed for Studies I and II; that is, SimStudent was given the same set of quiz problems each time the student administered a quiz. For Study III, the quiz problems were generated on the fly while keeping the *type* of problems intact. This means that although the numbers and variables letters were changed each time SimStudent took the quiz, the positive and negative signs were pre-






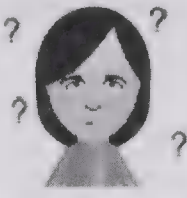

	Neutral	Thinking	Happy
i) Study I (Lucy)		N/A	N/A
ii) Study II (Stacy)			
iii) Study III (Tomodachi)			

Figure 2. SimStudent’s avatar image used in the three studies. There was only one static image used for Study I (Lucy). There are three facial expressions for Study II (Stacy): (a) neutral when waiting for the next problem to be entered, (b) thinking, and (c) happy after solving a problem. Study III (Tomodachi), which can be customized and uniquely named by each student, also has these three expressions. N/A = not applicable.

served. For example, $-3x + 5 = -7$ and $-2y + 1 = -10$ are considered to be isomorphic equations.

Self-Explanation

In Studies II and III, SimStudent had an ability to occasionally ask questions about things that students did, and these questions were intended to elicit students’ self-explanations (Matsuda, Cohen, et al., 2012).

SimStudent’s questions appear in the chat box at the bottom of the APLUS interface. SimStudent randomly selects a question from among a set of two to three questions that are relevant to each of three specific situations:

- 1. When the student inputs a *new problem* in the system, SimStudent asks why the student selected that problem or what that problem will help it learn.
- 2. If the student provides *negative feedback* on a step that SimStudent performed, SimStudent may or may not ask a question. If SimStudent has alternative actions to perform, it will not ask for an explanation. In cases in which SimStudent does solicit an explanation, it takes the last attempt that was made for the particular skill and asks why the step was incorrect, or how the situation is different from a previous step on which the same operation was used correctly.
- 3. SimStudent also asks questions after the student has provided a *hint* about transformation steps, not the results of the transformations (as the latter involves arithmetic calculation, and is thus often obvious). SimStudent will not ask a question at this point if it already asked about the student’s negative feedback on the same step.

The ways students input their response varies depending on the type of question. For a question on a demonstrated hint or new problem, there is a drop-down menu available with prewritten,

context-specific explanations. We hypothesized that such menu items would work as examples for students to learn (cf. Aleven & Koedinger, 2002). For questions about a demonstrated hint, the menu items use terminology such as *variable*, *constant*, and *coefficient* in a manner that reinforces their meanings. For questions about a new problem, the menu items include the key target concepts such as “It will help you learn how to deal with variables on both sides.” Even when selecting an answer from the drop-down menu, students can also edit the selected text with their own words. For questions about negative feedback, for example, “Why is (x) wrong?” students need to input their own answers. Figure 1b shows an example of a student’s response for SimStudent’s question about why “subtract 7” is wrong for first transformation (which, by the way, is an example of SimStudent making an error that students commonly make).

SimStudent waits for student input before continuing to the next step of the equation. After the student clicks the *submit* button, the answer appears in the chat box below SimStudent’s question. This explanation is also logged, but the answer does not affect SimStudent’s learning. If the student clicks the *submit* button without providing an explanation, the student has essentially ignored SimStudent’s question, and it will move on to the next step. In the classroom study, the students were not informed that they could skip the questions.

Overview of SimStudent’s Learning

The underlying machine-learning paradigm used for SimStudent is a technique called *programming by demonstration* (Lau & Weld, 1998) that generalizes positive and negative examples to generate a set of hypotheses using a given set of background knowledge sufficient to interpret (or “explain”) the examples. The positive and negative examples are provided by students as feedback and hints,

as described in the previous section. Affirmative feedback (i.e., “yes”) and hints become positive examples, whereas instances of negative feedback (i.e., “no”) become negative examples.

SimStudent generalizes from these positive and negative examples and generates a set of production rules that can reproduce all positive examples but no negative examples. Each production represents *where* to focus attention to know *when* and *how* to apply a particular skill. SimStudent uses hybrid AI techniques to learn the where, when, and how parts of a production rule. Providing technical details of the learning algorithm is beyond the scope of this article, but can be found elsewhere (Matsuda et al., 2007).

As mentioned earlier, one of the unique characteristics of SimStudent is its ability to learn skills incorrectly. We hypothesize that students learn incorrect skills by making inappropriate inductions from examples due to inappropriate background knowledge (Matsuda, Lee, Cohen, & Koedinger, 2009). Such incomplete background knowledge allows students to rely on shallow problem-solving features instead of deep domain principles.

As an example, suppose that a student is about to generalize an example of “subtracting 3 from both sides of $2x + 3 = 5$.” The student may recognize “+” in the left-hand side as the arithmetic operator instead of the sign of a term. As a consequence, the student may generalize this example to “subtract a number that follows an operator.” Students who perceive such a shallow feature would also be likely to subtract 4 from both sides of $3x - 4 = 6$ as well, which is one of the most frequently observed student errors (Booth & Koedinger, 2008).

To model this type of incorrect learning, we “weakened” SimStudent’s background knowledge by dropping the concept of an algebraic term in an expression and adding more perceptually grounded background knowledge, such as “get a number after an arithmetic operator.” In a prior study (Matsuda et al., 2009), we validated the cognitive fidelity of SimStudent’s learning by comparing SimStudent’s and human students’ learning. The study showed that SimStudent with “weak” prior knowledge learned skills incorrectly in a humanlike manner and generated humanlike errors when solving problems using the learned productions.

SimStudent applies learned productions to solve problems posed by a student, but the productions are not visible to the student. Therefore, the cognitive fidelity mentioned above could better facilitate tutor learning, because the student must identify, understand, and remediate SimStudent’s errors, which evoke or foster metacognitive tutoring skills and a deep understanding of the domain knowledge.

Method

Classroom Studies and Data Collection

The three in vivo studies were conducted as controlled randomized trials under the direct supervision of the Pittsburgh Science of Learning Center (LearnLab.org). Each study was conducted as a part of regular algebra classes. The studies used the same general format that involved 5 (Study I) or 6 (Studies II and III) days in the classroom. On the first day, all students took a pretest using an online test form (as described in the Measures section). After taking the pretest, students were randomly split into two groups and studied algebra equations using the assigned material for two

(Study I) or three (Studies II and III) class periods (one class period per day). All students then took an online posttest on the following day. Finally, all students took an online delayed test 2 weeks after the posttest.

Study I: Initial classroom trial. The primary goal of Study I was to evaluate the effectiveness of SimStudent (Matsuda et al., 2011). The version of APLUS and SimStudent used in Study I behaved exactly as described in the previous section and is called *Baseline* hereafter. Algebra I Cognitive Tutor (Ritter et al., 2007) was used for the control condition.

Study II: Self-explanation effect. In Study II, we focused on the self-explanation hypothesis, which conjectures that the tutor-learning effect is facilitated when the students are asked to explain and justify their tutoring decisions (Matsuda, Cohen, et al., 2012). To test this hypothesis, we compared SimStudent that did (the self-explanation condition) and did not (the baseline condition) ask questions.

Study III: Game show effect. For Study III, we compared the effect of learning by teaching SimStudent in APLUS with and without a Game Show feature (Matsuda, Yarzebinski, et al., 2012). In the Game Show, a pair of SimStudents, each tutored by a different student, compete by solving problems posed by the students who tutored them. This study was conducted to test the motivation hypothesis that conjectures that the more students are engaged in tutoring, the more tutor learning would be facilitated. The students in the Game Show condition were told to obtain the highest score in the Game Show, instead of having Tomodachi pass the quiz, which was the goal for the students in the non-Game Show condition.

Because the scope of this article does not include the motivation hypothesis, we do not discuss details of Study III here. However, we include Study III in the following analysis, because the control condition of Study III used the same version of SimStudent that Study II used for the self-explanation condition. Namely, the Study III SimStudent occasionally prompted students for explanations and justifications.

Participants

There were two schools involved in Study I. One school had 30 Algebra I (Grade 8) and 34 Algebra II (Grade 9) students, and the other school had 40 Algebra I (Grade 8) students. Study II involved one school with 160 Algebra I students in Grades 8, 9, and 10. Study III was conducted at the same school as Study II, and 141 Algebra I students in Grades 7 and 8 participated in Study III. To avoid a confounding factor of familiarity with the study, we excluded the ninth- and 10-grade students who were likely to have been included in Study II.

There were a significant number of absentees in each study. For the analysis in the following sections, we included only students who took all three (pre, post, and delayed) tests and participated in all classroom sessions. As a consequence, the following analyses contain 33 (32%), 81 (51%), and 69 (49%) of students for Study I, II, and III, respectively.

Measures

Outcome of tutee learning. To quantify tutee learning (i.e., SimStudent’s achievement), we use the number of quiz sections

that SimStudent passed, which differed in format among the three studies.

Outcome of tutor learning. Students' learning was measured with online tests that consisted of two parts—the *Procedural Skill Test* (PST) and the *Conceptual Knowledge Test* (CKT). The tests had three isomorphic versions that were counterbalanced for pre-, post-, and delayed tests. Two test items were considered isomorphic when they were of identical type, but included different letters and numbers. Equations were carefully varied so that two isomorphic equations shared the same properties in their solutions (e.g., whole number vs. fraction).

The PST had three types of test items: (a) six equation-solving items. Students were asked to show their work on a piece of paper; (b) twelve agree/disagree items to determine whether a given operation was a logical next step for a given equation; (c) five worked-out items to identify the incorrect step in a given incorrect solution (multiple choice) and explain why (free response). The CKT had two types of test items: (d) thirty-eight true/false items asking about basic algebra vocabulary to identify constants, variables, and like terms; (e) ten true/false items to determine whether two given expressions are equivalent.

For Studies II and III, the following changes were made on the online test: (a) Four additional one-step equations were added to the equation-solving items. (b) A "Not Sure" option was added for multiple-choice items to lower the chance of students making random guesses. Students were told that they would lose a point for an incorrect answer for multiple-choice questions, but there was no penalty for selecting "Not Sure."

The test items were graded as follows. For the equation-solving items, students received a score of 1 if their answer was correct and partial credit based on their written work if their answer was incorrect. For the multiple-choice items, students received a score of 1 for a correct answer, 0 for "Not Sure," and -1 for an incorrect answer.

Cognitive and Social Factors of Interest

APLUS automatically collects detailed data showing the interaction between students and SimStudent with additional narratives such as the response correctness. In the current analysis, we focus on the following variables:

1. The accuracy of students' *feedback* and *hints*. The accuracy of *response* is an aggregation of feedback and hints.
2. The likelihood of responding to SimStudent's hint request, which is the ratio of hints provided by a student to the total number of hints requested by SimStudent. Although students must answer SimStudent's hint request to proceed to the next step, they sometimes avoided answering by starting a new problem or giving a quiz.
3. The frequency of self-explanations submitted by students during tutoring.
4. The type of problems tutored. Although students were explicitly told that SimStudent must be able to solve equations with variables on both sides to pass the quiz, they needed to start with easier types to work up to the target difficulty.
5. The degree of repetition in selecting problems for tutoring. Students in our studies often used quiz problems during tutoring. As mentioned before, the problems in the quiz were fixed for Study II, but only the *type* of problems was fixed for Study III. To

avoid confusion, we shall use the term *problem* to mean the *exact* same problem for Study II and the same *type* of problem for Study III. The *problem repetition ratio* is then the ratio of the number of problems tutored more than once to the total number of problems tutored.

6. Time on task. The amount of time students spent tutoring problems and giving explanations to SimStudent. This time does not include the quiz or the resource usage.

7. Tutor's prior knowledge, that is, each student's PST and CKT pretest scores.

8. Tutee's learning outcome, that is, the number of quiz sections that SimStudent passed.

Results

This section is organized to answer the three major research questions mentioned in the introduction. We first show results about SimStudent's and students' learning outcomes addressing the first research question. We then show the correlations between tutor and tutee learning that answers the second research question. Finally, we show major findings obtained from the process data showing the cognitive and social factors that have significant influence on tutor and tutee learning.

Learning Outcomes

Because the three studies were conducted at different schools in different years, we first tested whether there was any population difference among the three studies. A one-way analysis of variance (ANOVA) was conducted with the independent variable of study (I, II, III) and the dependent variable of pretest score aggregated across two conditions. For both PST and CKT, the mean pretest score for Study I was significantly higher than Study II, which was significantly higher than Study III; for PST, $F(2, 180) = 23.58$, $p < .001$; for CKT, $F(2, 180) = 44.81$, $p < .001$. The difference in the pretest scores might reflect the age difference between the studies. Study III had the youngest student population.

Tutee-learning outcome: Performance on the quiz. Figure 3 shows the number of students whose SimStudent passed the quiz during the intervention. In Study I, none of the 18 students in the SimStudent condition managed to get their SimStudent to pass all four sections of the quiz. Only five students managed to pass quiz Section 1, and of those five, only one student passed quiz Section 3. For Study II, 36 out of 81 students managed to pass all four sections of the quiz within the allotted 3 days. Nearly all students (78 out of 81, i.e., 96%) passed at least Section 1. To our surprise, in the Study III baseline condition, we again observed that none of the students managed to have their SimStudent pass the quiz. Only 22 out of 40 (55%) students passed quiz Section 1. As mentioned earlier, Study III involved younger students and showed lower pretest score than the other two studies. The students in Study III might have less prepared for tutoring.

Tutor-learning outcome: Test scores. A summary of the test scores is shown in Table 1. The table shows mean scores for the pre-, post-, and delayed tests for all three studies. No condition difference on the pretest was found both for the PST and the CKT across the three studies. We thus conducted a 2×3 repeated measures ANOVA, with condition (study vs. control) as a between-subjects variable and test time (pre, post, and delayed) as

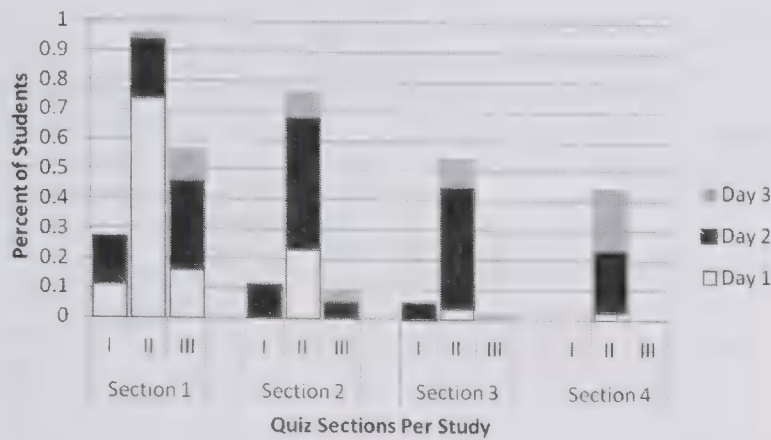


Figure 3. Number of students whose SimStudent passed the quiz. I, II, and III represents each study. Study I did not have a Day 3. The quiz had four sections with two equation problems per section, which were randomly ordered for Study I and ordered on the basis of difficulty level for Studies II and III. The quiz problems were fixed for Studies I and II. For Study III, the quiz problems were generated on the fly while keeping the type of problems intact.

a within-subjects repeated variable. The ANOVA was run on the PST and the CKT separately for each study.

For the PST, there was neither a main effect of condition nor an interaction effect between test time and condition. There was, however, an *iterative enhancement of the effect of tutor learning indicated as the gain on the test scores*. In Study I, the impact of test time was absent. In Study II, however, there was a main effect of test time, $F(2, 78) = 34.85, p < .001$. A further analysis showed that students' average test scores were significantly higher in the delayed test ($M = .68, SD = .27$) than both the pretest ($M = .54, SD = .26; p < .001, d = 0.53$) and the posttest ($M = .57, SD = .31; p < .001, d = 0.38$). The difference between the pre- and posttests was statistically indistinguishable. The reason for the higher delayed test scores in Study II might not be completely due to the study intervention. There were algebra classes between the post- and the delayed test (2 weeks apart) in which regular teachers continued teach-

ing equation solving. For Study III, again, there was a main effect of test time, $F(2, 66) = 8.81, p < .001$. Both posttest ($M = .45, SD = .20; p < .01, d = .35$) and delayed test ($M = .46, SD = .23; p < .001, d = .37$) were significantly higher than pretest ($M = .38, SD = .20$). The difference between the post- and delayed test was not statistically significant.

For the CKT, there was neither a main effect of test time nor a main effect of condition for all three studies.

Correlation Between Tutee and Tutor Learning

In Study III, tutee learning (the number of quiz sections that SimStudent passed) has a significant correlation with tutor learning (the normalized gain on the PST); $r(39) = .37, p < .05$. There was no significant correlation between tutee and tutor learning for Studies I and II.

Cognitive and Social Factors for Tutee and Tutor Learning

What affected tutee learning? It is surprising to observe that so many students failed to sufficiently tutor SimStudent to pass the quiz. To understand why, we conducted comparative analyses by splitting students into two groups based on the median quiz progress. For this analysis, we included students from both conditions in Study II ($N = 81$) and those in the control condition in Study III, in which the goal of tutoring was to have SimStudent pass the quiz ($N = 40$). Study I students were excluded from this analysis, because the order of quiz items in Study I was not compatible with Studies II and III.

Students were split into the *successful group* and the *unsuccessful group* using the median of the quiz section passed. For Study II, the split occurred at Section 3 (successful $n = 44$ vs. unsuccessful $n = 37$), whereas for Study III, the split occurred at Section 1 (successful $n = 21$ vs. unsuccessful $n = 18$). Within each study, we compared the two groups for a number of factors using independent samples t tests. Table 2 shows the results of this analysis.

Table 1
Test Scores Summary

Study	Pre-test		Post-test		Delayed-test	
I						
	CogTutor	Baseline	CogTutor	Baseline	CogTutor	Baseline
P	.67 (.24)	.74 (.19)	.73 (.20)	.76 (.21)	.65 (.25)	.72 (.17)
C	.55 (.13)	.62 (.15)	.57 (.10)	.63 (.16)	.52 (.19)	.58 (.13)
II						
	Baseline	SelfExpl	Baseline	SelfExpl	Baseline	SelfExpl
P	.54 (.27)	.52 (.26)	.57 (.29)	.57 (.33)	.68 (.25)	.68 (.30)
C	.29 (.23)	.29 (.27)	.32 (.24)	.33 (.27)	.30 (.25)	.35 (.26)
III						
	SelfExpl	Game Show	SelfExpl	Game Show	SelfExpl	Game Show
P	.34 (.19)	.44 (.19)	.41 (.21)	.49 (.19)	.43 (.23)	.50 (.23)
C	.13 (.20)	.19 (.20)	.16 (.19)	.21 (.20)	.15 (.19)	.22 (.20)

Note. CogTutor = cognitive tutor; Baseline = the baseline Artificial Peer Learning environment Using SimStudent (APLUS) and SimStudent; P = Procedural Skill Test; C = Conceptual Knowledge Test; SelfExpl = APLUS and SimStudent with self-explanation prompt; Game Show = APLUS and SimStudent with the Game Show feature. Each cell shows the mean, with the standard deviation in parentheses.

Table 2

Comparison Between Successful and Unsuccessful Groups Based on a Median Split for Number of Quiz Sections Passed

Factor	Study	Quiz performance		<i>t</i>	Cohen's <i>d</i>	<i>df</i>
		Successful	Unsuccessful			
Procedural Normalized Gain	II	.15 (.52)	.03 (.51)	-1.01	0.23	77 [†]
	III	.22 (.21)	-.01 (.29)	-2.80**	0.92	37 [†]
Correct Feedback	II	.86 (.07)	.83 (.08)	-1.52	0.34	79
	III	.78 (.06)	.68 (.14)	-2.80**	1.21	21 ^a
Correct Hint	II	.76 (.15)	.63 (.15)	-3.73***	0.83	79
	III	.56 (.13)	.30 (.25)	-3.95***	1.61	24 ^b
Disregarding Hint Requests	II	.44 (.16)	.52 (.20)	1.86	0.41	79
	III	.16 (.09)	.29 (.16)	3.14**	1.03	37 [†]
Repeating Problem Element	II	.13 (.13)	.23 (.19)	2.86**	0.72	62 ^c
	III	.38 (.12)	.46 (.19)	1.77	0.57	38

Note. Standard deviations appear in parentheses.

^a Levene's test indicated unequal variances, ($F = 10.07, p < .01$); *df* adjusted from 38 to 21. ^b Levene's test indicated unequal variances, ($F = 12.68, p < .001$); *df* adjusted from 38 to 24. ^c Levene's test indicated unequal variances, ($F = 9.48, p < .01$); *df* adjusted from 79 to 62.

** $p < .01$. *** $p < .001$. [†] Maximum *df* is 79 for Study II and 38 for Study III. Numbers marked with this symbol had 79-*N* or 38-*N* cases removed on the basis of a *z*-score outlier analysis of ± 3 .

First, the correctness of the student's *feedback* and *hints* had a notable influence on SimStudent's learning. For Study II, students in the successful group provided correct hints more often than students in the unsuccessful group. There was, however, no group difference in the accuracy of the feedback provided. For Study III, students in the successful group provided both correct hints and accurate feedback more often than students in the unsuccessful group.

Second, the likelihood of responding to SimStudent's hint request also has a notable difference. For Study III, the successful students responded to hint requests more often than unsuccessful students. The likelihood is, however, not significantly different between successful and unsuccessful groups for Study II.

Third, the problem repetition ratio was different. For Study II, the successful group tended to repeat the exact same problem less often than the unsuccessful group. For Study III, however, the difference was only marginal.

What affected tutor learning? In this analysis, we use the normalized gain of the PST from the pre- to the posttest as the measurement for the tutor learning. This analysis includes the same student data used in the tutee-learning analysis mentioned in the previous section.

First, the more the target problems were tutored (i.e., equations with variables on both sides), the more the *students* learned. This correlation was observed in both Study II, $r(79) = .26, p < .05$, and Study III, $r(39) = .36, p < .05$.

Second, the more the students gave self-explanations on the target problems, the more the students learned. Again, this correlation was observed in both Study II, $r(38) = .32, p < .05$, and Study III, $r(37) = .33, p < .05$.

Third, the more the students provided a correct tutoring responses (a combination of feedback and hint), the more the students learned, although this correlation was observed only in Study III, $r(38) = .31, p < .05$.

To our surprise, there was no correlation between the time on task and tutor learning in both studies: Study II, $r(81) = .09, p = .40$; Study III, $r(40) = .08, p = .61$.

Impact of Prior Knowledge for Tutor and Tutee Learning

We first show the impact of the tutee's prior knowledge on *tutor* learning. SimStudents for Study II (Stacy) and Study III (Tomodachi) were equally pretrained on more one-step equations than the SimStudent in Study I (Lucy). An independent samples *t* test confirmed that both Stacy and Tomodachi performed better on the first three tutoring problems than Lucy.

To see how the tutee's performance affected the tutor's performance, students' response accuracy was computed as a ratio of correct responses (i.e., feedback or hint) to all responses for each step in the first three tutored problems. On average, students in Studies II and III gave accurate responses more often than Study I students. Students in Studies II and III showed an average response accuracy of .76 ($SD = .21$), whereas students in Study I showed an average response accuracy of .57 ($SD = .26$). The difference is statistically significant, $t(134) = -3.49, p < .001, d = 0.60$.

One possible explanation for students' higher response accuracy in Studies II and III is that it is easier to recognize correct steps as correct than to identify incorrect steps as incorrect. Because Stacy and Tomodachi performed more steps correctly, the students in Studies II and III were able to correctly provide positive feedback more easily. When aggregated across all three studies, SimStudent's performance accuracy and students' response accuracy were actually highly correlated, $r(135) = .69, p < .001$.

Next, we analyzed the impact of the tutor's prior knowledge on tutor learning. The PST and CKT pretest were both predictive of students' posttest scores on the PST. A regression analysis with PST and CKT pretest scores as independent variables and the PST posttest score as a dependent variable revealed the following regression coefficients: $PST_Post = .70 \times PST_Pre + .12 \times CKT_Pre + .17$. An identical analysis was also conducted for the CKT posttest; a regression analysis with the PST and CKT pretest scores as independent variables and the CKT posttest score as a dependent variable revealed the following regression coefficients: $CKT_Post = .25 \times PST_Pre + .50 \times CKT_Pre + .05$.

Discussion

Results from the experiment provide four sets of important information to understand tutor learning. First, our data show that learning by teaching SimStudent is effective for learning procedural skills measured by the PST, as shown in Study III, but not for learning conceptual knowledge measured by the CKT.

Second, there is a significant correlation between tutee and tutor learning. Students tended to learn more when they tutored SimStudent correctly (i.e., with an accurate response) and appropriately (i.e., on appropriate problems with a sufficient amount of explanations).

Third, there were some notable differences in the way that the successful and the unsuccessful groups tutored SimStudent. Students in the unsuccessful group had trouble teaching SimStudent well, perhaps without even recognizing that they were not teaching appropriately. This manifested itself in students making many of the same errors, not properly responding to SimStudent's hint requests, and repeatedly teaching the same problem.

Fourth, both tutee and tutor's prior knowledge affected tutor learning. When the tutee had higher prior knowledge, the tutor tended to respond more accurately, which was further correlated with tutor learning. Our data also showed, however, that the tutor's prior competence both on conceptual and procedural knowledge was strongly predictive of tutor learning.

Finally, both SimStudent and APLUS have been iteratively improved from Study I to Study III, which may explain the gradual enhancement of the outcome. There was a population difference in the pretest score. Both for the PST and the CKT, students in Study I scored higher than the students in Study II, who outperformed the students in Study III. Yet, only Study III showed a significant gain in PST scores from pre- to posttest.

Tutor Help

Our findings show that learning by teaching does not happen automatically. Students need help to tutor SimStudent correctly and appropriately. Other research has also pointed out that students often do not correctly recognize their own misunderstandings (King, 1998). In our studies, students often unknowingly made inappropriate tutoring decisions and provided incorrect feedback and hints, which affected SimStudent's learning. These behaviors were negatively correlated with tutor learning.

One idea to provide such *tutor help* is to integrate a third agent (a *meta-tutor*) into the APLUS environment, a commonly used idea in the context of multiagent learning systems (e.g., Biswas et al., 2005; Vassileva, McCalla, & Greer, 2003). The meta-tutor oversees students' tutoring activities and provides them with just-in-time scaffolding.

The meta-tutor could provide students with both *cognitive* help regarding domain knowledge about how to solve problems and *metacognitive* help regarding proper tutoring methods. Some studies show that tutors can be trained to be a better tutor, which facilitates tutor learning (Ismail & Alexander, 2005; King et al., 1998). Other studies show the effect of the tutor help (Biswas et al., 2010; Walker et al., 2009), but none of them have explored the differing effects of cognitive and metacognitive help. It is therefore important to study how to implement cognitive and metacognitive help, how they foster tutor learning, and how well students learn tutoring skills from these different types of interactive support.

Another possibility is for the teachable agent itself to provide tutor help. For example, if a student poses the same problem (or the same type of problem) multiple times, then SimStudent could alert the tutor. The difference in the source of tutor help would have different social and affective impacts on the student. This might become particularly subtle when the student has established a different rapport with SimStudent and the meta-tutor. Studying the social factors of tutor help would therefore be important (Ogan, Finkelstein, Mayfield, D'Adamo, Matsuda, & Cassell, 2012).

The Effect of Self-Explanation for Tutor Learning

The current data show that the tutor-learning effect in APLUS is limited to procedural skills. Further studies will be needed to investigate the tutor-learning effect on conceptual knowledge. We hypothesized that self-explanations would facilitate learning conceptual knowledge, because good explanations contain conceptual justifications for algebraic operations. However, the students sometimes provided shallow responses (e.g., "Because you didn't add right") or irrelevant responses (e.g., "Because I just did"). The current version of SimStudent does not parse students' responses; instead, it simply proceeds to the next step. Empirical studies show that the tutee's questions have substantial influence on tutor learning (Roscoe & Chi, 2004). Thus, if SimStudent requested elaboration or further reflection on a given response, it may facilitate tutor learning. This kind of question is called a *reflective knowledge-building* question, and its effect has been well researched (Roscoe & Chi, 2007). Building such an intelligent teachable agent is therefore an important direction for future research (Carlson, Keiser, Matsuda, Koedinger, & Rose, 2012).

Learning by Teaching Versus Cognitive Tutoring

Our data show similarities and differences between learning by teaching and learning by cognitive tutoring (i.e., more direct instruction). The effect of self-explanation, for example, was evident for both styles of learning. Possession of prerequisite knowledge also has a notable influence on both styles of learning (Booth & Koedinger, 2008).

A notable difference between the two learning styles is the degree to which students can practice metacognitive skills. Our data show that the accuracy of tutoring responses, the frequency of self-explanations, and the type of problems tutored all positively correlate with tutor learning. To achieve successful learning, students must simultaneously monitor both their tutee's performances and their own. This double-edged monitoring requires more complicated metacognitive skills than solving problems alone in the context of cognitive tutoring.

There is also a difference in the timing of feedback. The feedback in the context of APLUS, that is, the system's reaction to the correctness of the student's tutoring activities, is delayed. The current version of APLUS does not provide students with any explicit feedback on their tutoring activities. Students later notice when they have made mistakes by reviewing the quiz summary or by observing SimStudent's undesired behaviors during tutoring. As an example of the second mistake, even when a student incorrectly demonstrated "subtract 4" for " $3x - 4 = 10$ " with a correct intention to isolate the " $3x$ " on the left-hand side, SimStudent might correctly suggest entering " $3x - 8$ " for the left-hand

side of the new equation, instead of “3x,” which is what the student expected to see.

The gap between the student’s expectations and SimStudent’s actual performance might motivate students to reflect on their tutoring actions. This is a kind of “intelligent novice” model of desired performance (Mathan & Koedinger, 2005) for tutor learning. Embedding the above-mentioned tutor help into the model of desired tutor performance might thus facilitate tutor learning. Observing the tutee’s performance is a distinctive form of learning from correct and incorrect examples available in learning by teaching.

Conclusion

Students learn by teaching others. Our data show that students learn by teaching primarily when they teach the target skills correctly and appropriately. The accuracy of students’ responses (i.e., feedback and hints), the quality of students’ explanations during tutoring, and the appropriateness of tutoring strategy (i.e., problem selection) all affected SimStudent’s learning outcome, which further affected students’ learning.

Students’ prior knowledge has a strong influence on tutor learning. If students are not well prepared to tutor, the benefits of tutor learning might be reduced. Alternatively, once students become domain experts and can solve problems fluently (hence become better teachers), the benefit of tutor learning might also decline. Tutor learning is essentially a paradoxical phenomenon whose mechanisms have yet to be fully elucidated.

Students make errors when teaching and get stuck when providing hints, both of which are detrimental for tutor learning. Providing more tutor help in the form of cognitive and/or meta-cognitive support may be critical to optimizing tutor learning.

The competence of the tutee also affects tutor learning as well. Carefully designing SimStudent’s learning ability and adaptively assigning an optimized SimStudent on the basis of the student’s competency would further provide us with insight into successful learning by teaching.

References

- Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, 26, 147–179. doi:10.1207/s15516709cog2602_1
- Annis, L. F. (1983). The processes and effects of peer tutoring. *Human Learning*, 2, 39–47.
- Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology*, 72, 593–604. doi:10.1037/0022-0663.72.5.593
- Bednarz, N., & Janvier, B. (1996). Emergence and development of algebra as a problem solving tool: Continuities and discontinuities with arithmetic. In N. Bednarz, C. Kieran, & L. Lee (Eds.), *Approaches to algebra* (pp. 115–136). Dordrecht, the Netherlands: Kluwer.
- Biswas, G., Jeong, H., Kinnebrew, J. S., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning*, 5, 123–152.
- Biswas, G., Leelawong, K., Belyne, K., Viswanath, K., Vye, N. J., Schwartz, D. L., & Davis, J. (2004). Incorporating self-regulated learning techniques into learning by teaching environments. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 120–125). Mahwah, NJ: Erlbaum.
- Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & The Teachable Agents Group at Vanderbilt. (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence: An International Journal*, 19, 363–392. doi:10.1080/08839510590910200
- Booth, J. L., & Koedinger, K. R. (2008). Key misconceptions in algebraic problem solving. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 571–576). Austin, TX: Cognitive Science Society.
- Carlson, R., Keiser, V., Matsuda, N., Koedinger, K. R., & Rose, C. (2012). Building a conversational SimStudent. In S. Cerri & W. Clancey (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 563–569). Heidelberg, Berlin: Springer-Verlag.
- Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25, 471–533. doi:10.1207/s15516709cog2504_1
- Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research*, 64, 1–35. doi:10.2307/1170744
- Cohen, P. A., Kulik, J. A., & Kulik, C-L. C. (1982). Education outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19, 237–248.
- Cook, S. B., Scruggs, T., Mastropieri, M., & Casto, G. (1986). Handicapped students as tutors. *Journal of Special Education*, 19, 483–492. doi:10.1177/002246698501900410
- Devin-Sheehan, L., Feldman, R. S., & Allen, V. L. (1976). Research on children tutoring children: A critical review. *Review of Educational Research*, 46, 355–385. doi:10.2307/1170008
- Filloy, E., & Rojano, T. (1989). Solving equations: The transition from arithmetic to algebra. *For the Learning of Mathematics*, 9, 19–25.
- Gartner, A., Kohler, M., & Riessman, F. (1971). *Children teach children: Learning by teaching*. New York, NY: Harper & Row.
- Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9, 495–522. doi:10.1002/acp.2350090604
- Greenwood, C. R., Delquadri, J. C., & Hall, R. V. (1989). Longitudinal effects of classwide peer tutoring. *Journal of Educational Psychology*, 81, 371–383. doi:10.1037/0022-0663.81.3.371
- Ismail, H. N., & Alexander, J. M. (2005). Learning within scripted and nonscripted peer-tutoring sessions: The Malaysian context. *The Journal of Educational Research*, 99, 67–77. doi:10.3200/JOER.99.2.67-77
- Jacobson, J., Thrope, L., Fisher, D., Lapp, D., Frey, N., & Flood, J. (2001). Cross-age tutoring: A literacy improvement approach for struggling adolescent readers. *Journal of Adolescent & Adult Literacy*, 44, 528–536.
- Kieran, C. (1992). The learning and teaching of school algebra. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 390–419). New York, NY: MacMillan.
- King, A. (1998). Transactive peer tutoring: Distributing cognition and metacognition. *Educational Psychology Review*, 10, 57–74. doi:10.1023/A:1022858115001
- King, A., Staffieri, A., & Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology*, 90, 134–152. doi:10.1037/0022-0663.90.1.134
- Lau, T. A., & Weld, D. S. (1998). Programming by demonstration: An inductive learning formulation. *Proceedings of the 4th International Conference on Intelligent User Interfaces* (pp. 145–152). New York, NY: ACM Press.
- Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty’s Brain system. *International Journal of Artificial Intelligence in Education*, 18, 181–208.
- Li, N., Matsuda, N., Cohen, W. W., & Koedinger, K. R. (2011). A machine learning approach for automatic student model discovery. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, & J. Stamper (Eds.), *Proceedings of the International Conference on Educational*

- Data Mining* (pp. 31–40). Eindhoven, the Netherlands: International Educational Data Mining Society.
- Lincevski, L., & Herscovics, N. (1996). Crossing the cognitive gap between arithmetic and algebra: Operating on the unknown in the context of equations. *Educational Studies in Mathematics*, 30, 39–65. doi:10.1007/BF00163752
- Mastropieri, M. A., Spencer, V. G., Scruggs, T. E., & Talbott, E. (2000). Students with disabilities as tutors: An updated research synthesis. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Advances in learning and behavioral disabilities* (Vol. 14, pp. 247–279). Stamford, CT: JAI Press.
- Mathan, S. A., & Koedinger, K. R. (2005). Fostering the intelligent novice: Learning from errors with metacognitive tutoring. *Educational Psychologist*, 40, 257–265. doi:10.1207/s15326985ep4004_7
- Mathes, P. G., & Fuchs, L. S. (1994). The efficacy of peer tutoring in reading for students with mild disabilities: A best-evidence synthesis. *School Psychology Review*, 23, 59–80.
- Matsuda, N., Cohen, W. W., Koedinger, K. R., Keiser, V., Raizada, R., Yarzebinski, E., & Stylianides, G. J. (2012). Studying the effect of tutor learning using a teachable agent that asks the student tutor for explanations. In M. Sugimoto, V. Aleven, Y. S. Chee, & B. F. Manjon (Eds.), *Proceedings of the International Conference on Digital Game and Intelligent Toy Enhanced Learning (DIGTEL 2012)* (pp. 25–32). Los Alamitos, CA: IEEE Computer Society.
- Matsuda, N., Cohen, W. W., Sewall, J., Lacerda, G., & Koedinger, K. R. (2007). Evaluating a simulated student using real students data for training and testing. In C. Conati, K. McCoy, & G. Paliouras (Eds.), *Proceedings of the International Conference on User Modeling (LNAI 4511)* (pp. 107–116). Berlin, Germany: Springer.
- Matsuda, N., Lee, A., Cohen, W. W., & Koedinger, K. R. (2009). A computational model of how learner errors arise from weak prior knowledge. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 1288–1293). Austin, TX: Cognitive Science Society.
- Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., Stylianides, G. J., Cohen, W. W., & Koedinger, K. R. (2011). Learning by teaching SimStudent—An initial classroom baseline study comparing with cognitive tutor. In G. Biswas & S. Bull (Eds.), *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 213–221). Berlin, Germany: Springer.
- Matsuda, N., Yarzebinski, E., Keiser, V., Raizada, R., Stylianides, G. J., & Koedinger, K. R. (2012). Motivational factors for learning by teaching: The effect of a competitive game show in a virtual peer-learning environment. In S. Cerri & W. Clancey (Eds.), *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 101–111). Heidelberg, Germany: Springer-Verlag.
- Michie, D., Paterson, A., & Hayes, J. E. (1989). *Learning by teaching. Proceedings of the Second Scandinavian Conference on Artificial Intelligence* (pp. 413–436). Amsterdam, the Netherlands: IOS.
- Ogan, A., Finkelstein, S., Mayfield, E., D'Adamo, C., Matsuda, N., & Cassell, J. (2012). “Oh, dear Stacy!” Social interaction, elaboration, and learning with teachable agents. In *Proceedings of the International Conference on Computer–Human Interaction (CHI2012)* (pp. 39–48). Austin, TX: ACM.
- Ohlsson, S. (2008). Computational models of skill acquisition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 359–395). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511816772.017
- Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1, 117–175. doi:10.1207/s1532690xci0102_1
- Pareto, L., Arvemo, T., Dahl, Y., Haake, M., & Gulz, A. (2011). A teachable-agent arithmetic game's effects on mathematics understanding, attitude and self-efficacy. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 247–255). Heidelberg, Germany: Springer.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1, 81–106. doi:10.1007/BF00116251
- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14, 249–255. doi:10.3758/BF03194060
- Robinson, D., Schofield, J., & Steers-Wentzell, K. (2005). Peer and cross-age tutoring in math: Outcomes and their design implications. *Educational Psychology Review*, 17, 327–362. doi:10.1007/s10648-005-8137-2
- Rohrbeck, C. A., Ginsburg-Block, M., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. *Journal of Educational Psychology*, 95, 240–257. doi:10.1037/0022-0663.95.2.240
- Roscoe, R. D., & Chi, M. T. H. (2004). The influence of the tutee in learning by peer tutoring. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 1179–1184). Mahwah, NJ: Erlbaum.
- Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, 77, 534–574. doi:10.3102/0034654307309920
- Sharpley, A. M., Irvine, J. W., & Sharpley, C. F. (1983). An examination of the effectiveness of a cross-age tutoring program in mathematics for elementary school children. *American Educational Research Journal*, 20, 103–111.
- Topping, K. J. (1996). The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education*, 32, 321–345. doi:10.1007/BF00138870
- Ur, S., & VanLehn, K. (1995). STEPS: A Simulated, Tutorable Physics Student. *Journal of Artificial Intelligence in Education*, 6, 405–437.
- VanLehn, K., Ohlsson, S., & Nason, R. (1994). Applications of simulated students: An exploration. *Journal of Artificial Intelligence in Education*, 5, 135–175.
- Vassileva, J., McCalla, G., & Greer, J. (2003). Multi-agent multi-user modeling in I-Help. *User Modeling and User-Adapted Interaction*, 13, 179–210. doi:10.1023/A:1024072706526
- Walker, E., Rummel, N., & Koedinger, K. R. (2009). CTRL: A research framework for providing adaptive collaborative learning support. *User Modeling and User-Adapted Interaction*, 19, 387–431. doi:10.1007/s11257-009-9069-1
- Webb, N. M., & Mastergeorge, A. (2003). Promoting effective helping behavior in peer-directed groups. *International Journal of Educational Research*, 39, 73–97. doi:10.1016/S0883-0355(03)00074-0

Received December 16, 2011

Revision received November 16, 2012

Accepted December 10, 2012 ■

Gendered Socialization With an Embodied Agent: Creating a Social and Affable Mathematics Learning Environment for Middle-Grade Females

Yanghee Kim
Utah State University

Jae Hoon Lim
University of North Carolina at Charlotte

This study examined whether or not embodied-agent-based learning would help middle-grade females have more positive mathematics learning experiences. The study used an explanatory mixed methods research design. First, a classroom-based experiment was conducted with one hundred twenty 9th graders learning introductory algebra (53% male and 47% female; 51% Caucasian and 49% Latino). The results revealed that learner gender was a significant factor in the learners' evaluations of their agent ($\eta^2 = .07$), the learners' task-specific attitudes ($\eta^2 = .05$), and their task-specific self-efficacy ($\eta^2 = .06$). In-depth interviews were then conducted with 22 students selected from the experiment participants. The interviews revealed that Latina and Caucasian females built a different type of relationship with their agent and reported more positive learning experiences as compared with Caucasian males. The females' favorable view of the agent-based learning was largely influenced by their everyday classroom experiences, implying that students' learning experience in real and virtual spaces was interconnected.

Keywords: embodied agents, interactive learning environments, equity in mathematics education, human-computer interaction

A recent analysis of the National Assessment of Educational Progress data reported that the achievement gap between Caucasians and ethnic minority students (e.g., African Americans and Latinos) in mathematics achievement has become stagnant during the last two decades (Vanneman, Hamilton, Anderson, & Rahman, 2009). Female students, despite their improved achievement in mathematics (Lindberg, Hyde, Petersen, & Linn, 2010), still report lower interest and lower self-confidence in mathematics as compared with males (Jacobs, Davis-Kean, Bleeker, Eccles, & Malanchuk, 2005). These underrepresented groups of students often “disidentify” themselves with mathematics learning (Steele, Spencer, & Aronson, 2002) and, as a result, are more likely to avoid taking advanced mathematics classes (Steffens, Jelenec, & Noack, 2010). Acknowledging the urgency in resolving these problems, the National Science Board (2010) has declared its commitment to equity and diversity as a focal area for developing the next generation of science, technology, engineering, and mathematics (STEM) innovators.

A variety of social, cultural, and economic factors might lead to the equity issues. However, gender and ethnic inequity in

mathematics education is often attributed to the unsupportive learning context in schools (Moody, 2004) and undesirable social influences such as stereotyping (Steele et al., 2002). Females and ethnic minorities often lack the instructional support that might motivate them to engage and succeed in the area. This lack of support, coupled with social stereotyping, leads them to hold a negative view of mathematics and to doubt their capability to succeed.

Reshaping the school context and social influences might be a long societal process, requiring synergistic endeavors by a multitude of individuals and institutions. Nonetheless, advanced learning technology might design supportive learning contexts that help close these motivational and achievement gaps. One such technology that uses animated digital characters (called *embodied interface agents*) promises to augment the bandwidth of a learner's interactions with computers (Bailenson et al., 2008) and to add social richness to the interactions (Iacobelli & Cassell, 2007). Many females' and ethnic minorities' learning styles favor active and multifaceted interactions (Sciarra & Seirup, 2008); connectedness and relationships are characteristic of their learning process (Crosnoe et al., 2010). If designed carefully, agent-based learning might be able to create a favorable learning context for these students, accommodating their learning styles and characteristics.

This study was conducted to examine this expectation that the females' and minorities' affect and learning would improve in a more social and affable agent-based environment. The study, consisting of a classroom experiment and following in-depth interviews, investigated how middle-grade students learning introductory algebra would react to an agent and whether the reactions would differ by the students' gender and ethnicity.

This article was published Online First September 9, 2013.

Yanghee Kim, Department of Instructional Technology and Learning Sciences, Utah State University; Jae Hoon Lim, Department of Educational Leadership, University of North Carolina at Charlotte.

This research was supported in part by National Science Foundation Grant HRD-0522634.

Correspondence concerning this article should be addressed to Yanghee Kim, Department of Instructional Technology and Learning Sciences, Utah State University, 2830 Old Main Hill, Logan, Utah 84321. E-mail: yanghee.kim@usu.edu

Theoretical Background

Sociocultural Aspect of Mathematics Learning

The sociocultural context of learning plays a significant role in shaping students' motivation, learning behaviors, and academic outcomes in schools. The learning process is not merely a cognitive restructuring within an individual mind. It is a social and cultural process in which multiple facets of human development (e.g., identity and emotion) are intertwined with social, cultural, and historical forces (Nasir, Rosebery, Warren, & Lee, 2006). For example, a student's "sense of belonging" in school positively correlates with her strong and clear identification with the goal of schooling, which ultimately leads her to full, active participation in all aspects of the learning process (Freeman, Anderman, & Jensen, 2007).

The mathematics learning context and its social and instructional dynamics play a critical role in motivating all students to learn and excel in mathematics. However, the context and dynamics seem to have even more critical influence on traditionally underrepresented groups of students (Geist, 2010). Feminist scholars argue that females' unique way of learning is not best supported by the traditional mathematics classrooms (Boaler, 2002). Females are "connected knowers" and tend to rely on interpersonal relationships and commonality of experience when they approach a new idea or knowledge (Belenky, Clinchy, Golberger, & Tarule, 1997). Supportive relationships with instructional authority and peers might be critical for many females' intellectual pursuit of mathematics and perseverance in the area (Crosnoe, Riegler-Crumb, Field, Frank, & Muller, 2008). However, the mathematics education community has a long tradition that views mathematics learning as a depersonalized activity disconnected from other aspects of students' everyday lives (Cobb & Yackel, 1998). This assumption about mathematics learning disregards the typical style of female learning. Not surprisingly, many females experience higher anxiety and discomfort in mathematics classes than boys (Geist, 2010). These females report lower interest and self-efficacy even when their performances are equal to or better than boys' during early school years (Lindberg et al., 2010). As a result, the females avoid taking advanced mathematics courses in high school (Steffens et al., 2010).

A similar phenomenon is observed among many Latino students. Three types of engagement influence Latinos' achievement in mathematics: cognitive, emotional, and behavioral. Latinos show a higher level of engagement in mathematics when they are asked to work with peers than when asked to work alone (Uekawa, Borman, & Lee, 2007). They are more likely to use a participatory communication style, which requires active response from the audience, such as verbal encouragement or even physical movement during speech (Gay, 2000). This form of communication is not readily accepted in conventional mathematics classrooms. Rather, it is often viewed as a disruptive behavior or, at best, an attitude less effective for learning (Neal, McCray, Webb-Johnson, & Bridgest, 2003). Not surprisingly, Latino students experience a higher level of mathematics anxiety than Caucasian students; Latinas' anxiety tends to be even worse than that of their male counterparts (Willig, Harnisch, Hill, & Maehr, 1983).

Embodied Agents to Create a Social and Affable Context

Although computers are often regarded merely as a tool to perform tasks, computer users actually tend to expect computers to be like social entities (Lee & Nass, 2003). In response to animated digital characters, users build humanlike relationships with the character (Bickmore, 2003); college students expect a digital character acting as a tutor to have a nice personality as well as content expertise (Kim, 2007). Furthermore, just as girls' and boys' preferences for instructional content, activities, and methods are differentiated in classrooms, so are their reactions to the features in computer-based learning (Kinzie & Joseph, 2008). Females' inclinations toward interactions and relationship building in classrooms are consistently demonstrated in computer-based environments. For instance, girls like interactive and dynamic hints from the computer more than do boys (Arroyo, Murray, Woolf, & Beal, 2003).

Researchers in educational technology have explored the use of embodied interface agents in various theoretical and practical frameworks, for example, in the framework of computer-supported collaboration (White, Shimoda, & Frederiksen, 1999), or as a way to render a sense of social presence (Graesser, Chipman, Haynes, & Olney, 2005; Moreno & Flowerday, 2006). Embodied agents even seem to play a persuasive role in shaping viewpoints, attitudes, and behaviors. One experiment revealed that an agent's pedagogical perspectives were successfully projected into college students' own pedagogical perspectives. Preservice teachers who worked with an agent who took a constructivist perspective adopted the constructive perspective after their interactions with the agent, whereas those who worked with an agent taking an objectivist perspective adopted the objectivist perspective (Baylor, 2002). In another study, middle-school students who had received instructions from an agent reported lower levels of perceived difficulty than did the students who had received textual information without an agent (Atkinson, 2002). Also, when kindergarten children played with the virtual peer *Sam*, they listened to *Sam*'s stories very carefully and, afterward, mimicked *Sam*'s linguistic styles (Ryokai, Vaucelle, & Cassell, 2003).

Traditionally, human one-on-one tutoring has been considered the best form of instruction because it increased learning by two standard deviations as compared with the group instruction in a classroom (Bloom, 1984). Researchers in computer-based tutoring have strived to approximate the effect of human tutoring. As a result, successful tutoring systems were able to increase learning by one standard deviation higher than the control groups (Graesser et al., 2005; Koedinger & Anderson, 1997). This success has raised the inquiry into how we can further make up the missing one standard deviation effect through our design. Stone (1998) identified three components of scaffolding—perceptual, cognitive, and affective—as necessary for effective learning and motivation. Many conventional tutoring environments, however, have focused on assisting learners in only the cognitive processes of learning. They often neglected to implement perceptual and affective aspects of scaffolding. Recently, researchers in educational technology have come to better understand the integral role of human cognition and affect in the learning process, and have made efforts to equip tutoring environments with affective capabilities

(D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008; du Boulay et al., 2010).

These trends in embodied agents and tutoring systems provides a further line of inquiry into the potential of embodied agent technology for addressing the equity issues in mathematics education. When teachers and parents simply presented the facts about the nonexistence of gender difference in mathematics learning, adolescent females became able to resist negative stereotypes concerning girls and mathematics (Jacobs et al., 2005). By presenting similar messages in the course of instructional guidance, an embodied agent tutor might inculcate positive attitudes toward mathematics learning and improve females' self-efficacy beliefs. If this expectation turns out to be true, embodied agent technology will be able to expand the functionality of conventional tutoring systems, which have typically assumed a motivated learner instead of generating motivation (du Boulay et al., 2010).

In this study, we conducted two phases of empirical inquiry. The first phase was an in vivo experiment, in which quantitative data were collected in natural classrooms. In the second phase, in-depth interviews were conducted to better understand the nature of students' learning experiences with their agent. The guiding research question was: Will middle-grade females' and Latinos' reactions to an embodied agent be qualitatively different from Caucasian males' reactions?

Classroom Experiment

Hypothesis

In this classroom-based experiment, we investigated whether or not learner gender and ethnicity would influence learners' evaluations of their agent, mathematics attitudes, mathematics self-efficacy, and learning gains. We tested four hypotheses: (a) Females and Latino students would evaluate their agent more positively than Caucasian males; (b) females' and Latinos' attitudes toward learning mathematics from their agent would be more positive than Caucasian males' attitudes; (c) females' and Latinos' self-efficacy in learning mathematics from their agent would be higher than Caucasian males' self-efficacy; and (d) females and Latinos would increase their learning similar to Caucasian males after the intervention. In addition, if an embodied agent would have a positive influence on females and Latinos, theoretically, Latinas would be the group benefiting most from the agent; Caucasian males would benefit least. The two groups were compared in each of the dependent measures.

Method

Participants. Participants were one hundred twenty 9th graders enrolled in Algebra I classes in two inner-city high schools in a mountain-west state in the United States. Sixty-four students were male (53%) and fifty-six were female (47%). Sixty-one students were Caucasian (51%) and fifty-nine Latino (49%). In the participating school districts, students were able to start taking Algebra I in the seventh grade and required to complete it by the ninth grade. Thus, the participants who had delayed the course until required were assumed to be less interested in mathematics than the rest of the ninth graders in the schools. The average age of the participants was 15.93 ($SD = 0.87$).

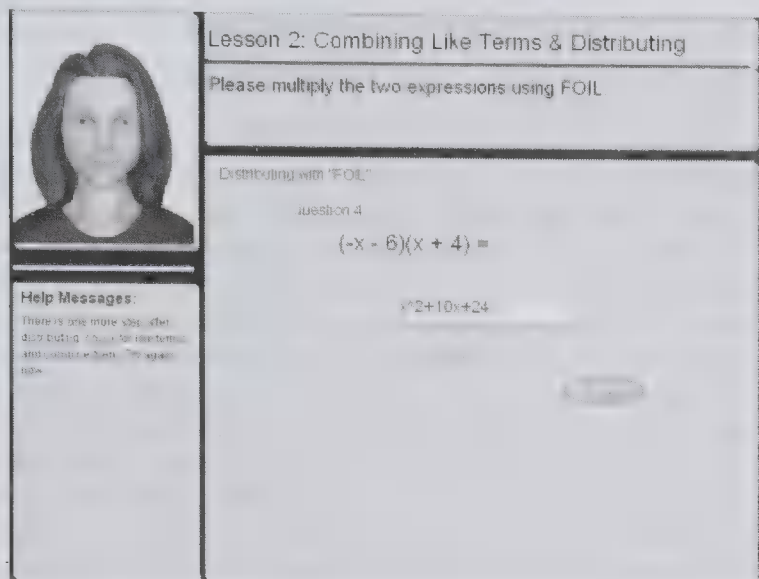
Intervention: An agent-based algebra-learning environment.

The intervention was two computer-based lessons, integrated with an embodied agent. The lessons were designed as supplemental materials to classroom learning, in which a learner reviewed the concepts individually that he or she had learned from the teacher and practiced solving problems to master the concepts. The agent, designed as a tutor, presented curriculum-related information and feedback and also verbally encouraged the learner to sustain in the task. The learning environment was self-contained, within which the learner typed in demographic information to log in, performed the learning task, and took pre- and posttests.

Curricular content. Following the *Principles and Standards of the National Council of the Teachers of Mathematics* (<http://www.nctm.org>), the curriculum was developed in collaboration with the algebra teachers in the participating schools, addressing their classroom needs. The two lessons, each taking one class period (approximately 50 min), dealt with combining like terms and distributive properties (Lesson 1) and graphing linear equations using slope and y-intercept (Lesson 2). The lessons consisted of four to five sections, each section including two phases: (a) Review of Concepts and (b) Problem Practice. In the Review, the agent presented brief overviews of key concepts and examples. In the Problem Practice, a learner solved problems one at a time by way of drill-and-practice, listening to the agent's feedback. The lessons were prescribed so that every learner could be exposed to all overviews and solve the same number of problems. The teachers helped identify the errors that students typically made in the classroom and helped write corrective feedback messages. Figure 1 presents example screens of the lesson environment.

Agent design. The design goal for the embodied agent named *Chris* was to simulate the instructional, social, and empathetic roles that might be played by an effective human tutor. We achieved the goal by including three features in agent design: (a) personalized instructions, (b) social and empathic rhetoric, and (c) peerlike image and voice. Regarding personalized instructions, while a learner worked individually at his or her own pace, *Chris* used the personal pronoun *we* in its explanations and feedback, emphasizing "a sense of togetherness." The problem for the learner to solve was not his or hers but "our problem." For social and empathic rhetoric, *Chris* used two types of messages in addition to curricular overviews and feedback: motivational and persuasive. Motivational messages were words of praise and verbal encouragement presented when the learner made a mistake. Persuasive messages were statements about the benefits or advantages of doing mathematics well. The persuasive messages were integrated into the introductions to new sections and subsections so that every learner would hear persuasive statements. To promote agent-learner affinity, the messages adopted the teenagers' style of speech. Two high school students translated the messages developed by the design team into such teen-friendly speech. The Appendix presents examples of the agent messages. Lastly, we used peerlike image and voice to increase a sense of affinity. To control for the confounding effect by learners' biases toward agent gender or ethnicity (Kim & Wei, 2011), we developed four versions of an agent to match the students' gender and ethnicity. One of the four agents was randomly assigned to a student. Also, to control for the confounding effect by agent appearance (Gulz & Haake, 2006), we morphed the four versions from one base image. Following that, we validated the agent images with another group

Lesson 1: Combining Like Terms



Lesson 2: Graphing Linear Equations

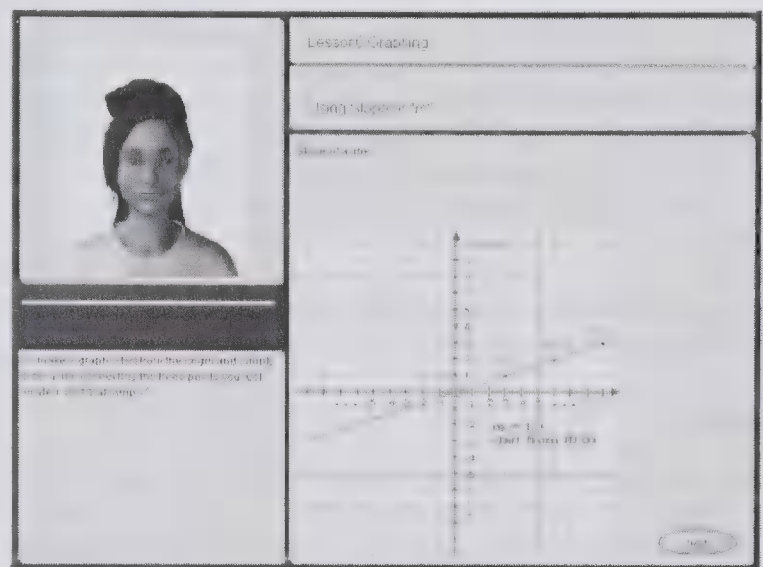


Figure 1. The example screens of the agent-based learning environment.

of 200 high school students, which confirmed that the images looked exactly as intended. Agent voices were recorded by four adolescent voice actors, in consideration of the friendliness of the human voice as compared with a synthesized one (Mayer, Sobko, & Mautone, 2003). Lastly, we added facial expressions, eye gaze, and head nodding to make the agents look more natural and believable.

Variables and measures. Independent variables included learner gender and ethnicity, each having two levels: male Caucasians (35), female Caucasians (26), Latinos (29), and Latinas (30). Dependent measures were learners' evaluations of an agent, mathematics attitudes, mathematics self-efficacy, and learning gains.

Learners' evaluations of an agent. Learners' evaluations of their agent were measured with a 17-item questionnaire using a Likert scale ranging from 1 (*Strongly disagree*) to 7 (*Strongly agree*). The items asked whether the agent was friendly and helpful for learning and whether the learner desired to work with the agent again (e.g., "Chris was friendly," "Chris was easy to understand," and "I'd like to learn from Chris again"). Interitem reliability was evaluated as $\alpha = .96$.

Mathematics attitudes. *Mathematics attitudes* were defined as learners' overall evaluative responses to learning mathematics (Petty, DeSteno, & Rucker, 2001). Pre- and posttest items, scaled from 1 (*Strongly disagree*) to 7 (*Strongly agree*), were derived from the Attitudes Toward Mathematics Inventory (<http://www.rapidintellect.com/AEQweb/cho253441.htm>). The five-item pretest measured learners' general attitudes toward learning mathematics (e.g., "In general, I like learning math"). The pretest was used as a covariate in the analysis; the interitem reliability evaluated with coefficient $\alpha = .80$. The posttest included two categories of attitudes. One category measured learners' general attitudes (same as the pretest); the other measured learners' attitudes specifically toward learning mathematics from the agent in the lessons (two items), for example, "I liked solving math problems with Chris in this lesson." Posttest interitem reliability was evaluated as $\alpha = .84$.

Mathematics self-efficacy. *Mathematics self-efficacy* was defined as learners' beliefs in their capability to successfully learn mathematics (Bandura, 1997). Following Bandura's (2006) guidelines, pre- and posttest items were developed and ranged from 1 (*Strongly disagree*) to 7 (*Strongly agree*). The five-item pretest measured the learners' general self-efficacy beliefs in learning mathematics (e.g., "In general, I am confident in learning math"). The pretest was used as a covariate; interitem reliability was evaluated as $\alpha = .84$. The posttest included two categories of self-efficacy. One category measured learners' general self-efficacy (same as the pretest); the other category measured their self-efficacy specifically in learning mathematics from the agent (four items), for example, "I was confident in solving problems with Chris in this lesson." Posttest interitem reliability was evaluated as $\alpha = .86$.

Learning gains. Learning was measured with a pretest and an immediate posttest. After logging into the system, the learners solved 16 problems; at the end of the lesson, they solved another set of 16 equivalent problems. For example, one item in pretest asked the learners to distribute the expression $3a(x + y)$; the matching posttest item asked to distribute the expression $5x(a + b)$. The items were presented one after another; the format was similar to Figure 1, Question 4 on the left, without agent presence. Students used scratch paper and pencil to solve a problem and typed in their answers in a blank. Each item was scored correct (1) or incorrect (0), with the maximum score of 16 and no partial scores.

Procedure. We implemented the experiment as regular activities in the classroom (using 34 laptop computers) on 2 consecutive days, one lesson per day. On Day 1, students were given a brief introduction about the lesson and interface and then asked to put on headphones. They entered demographic information to log onto the lesson. Upon login, they took pretests. Following that, one of the four agents (differing in gender and ethnicity) was randomly assigned to a student. Students performed the learning task, listening to Chris' overviews and feedback. On Day 2, students were

assigned to the same agent and performed the learning task in the same manner. Lastly, they took posttests without Chris.

Design and analysis. A 2×2 factorial design was used, in which both learner gender and ethnicity had two levels. To analyze learners' evaluations of an agent, a two-way analysis of variance (ANOVA) was conducted. To analyze attitudes and self-efficacy (each having two subcategories), 2 two-way multivariate analyses of covariance (MANCOVAs) were conducted, respectively, with a pretest set as a covariate to control for the group difference in the pretest. To analyze learning, a two-way repeated analysis of covariance (ANCOVA) was conducted, with a pretest set as a covariate. The significance level was set at $\alpha < .05$.

Results

A preliminary analysis of the data was conducted to ensure that the assumptions of the parametric statistics were met. Visual examination of scatterplots supported the assumption of normality and revealed linear relationships. Levene's test was conducted to test the equality of error variance for each ANOVA procedure; Box's test was conducted to test the equality of covariance for each MANCOVA procedure. These tests did not reveal any significant problems with the equality of error variance and covariance. Table 1 presents the means and standard deviations for learners' evaluations of an agent, mathematics attitudes, and mathematics self-efficacy.

Gendered and ethnicity-based positivity of agent evaluations. The two-way ANOVA indicated a significant main effect of learner gender, $F(1, 116) = 8.22, p = .005, \eta^2 = .07$. The females evaluated their agent significantly more positively than did the males. Also, there was a significant main effect of learner ethnicity, $F(1, 116) = 22.87, p = .000, \eta^2 = .17$. The Latinos evaluated their agent significantly more positively than did the Caucasians. A planned two-independent group t test was further conducted to compare Latinas with Caucasian males. The results revealed that the Latinas evaluated their agent significantly more positively than did the Caucasian males ($t = -5.54, p = .000, d = -1.39$).

Gendered inflection of attitudes. The two-way MANCOVA revealed a significant main effect of learner gender (Wilks's $\Lambda = .95$), $F(2, 114) = 2.95, p = .046$, partial $\eta^2 = .05$. Given the overall significance, a univariate analysis was further conducted to examine the contribution of each category of attitudes to the overall significance. There was a significant main effect of learner gender on the attitudes specifically toward learning mathematics in the agent-based lessons, $F(1, 115) = 4.82, p = .030, \eta^2 = .04$. The

females showed significantly more positive attitudes than did the males. A planned contrast between Latinas and Caucasian males revealed a similar pattern that the Latinas showed significantly more positive attitudes than the Caucasian males ($t = -2.42, p = .019, d = -0.6$).

Gendered enhancement of self-efficacy. The two-way MANCOVA revealed neither main effect nor interaction effect of learner gender and learner ethnicity on learners' mathematics self-efficacy ($p = .783$). Nonetheless, the goal of the learning environment was to help students build their confidence in mathematics learning; we inquired about any group difference in the improvement of their self-efficacy after the intervention. A two-way repeated ANOVA was conducted to examine changes in learner self-efficacy from pretest to posttest. Because the number of the items in the two tests was not matched, the posttest scores were statistically converted to match the pretest scores. There was a significant interaction effect of the within-subject factor (time) and learner gender, $F(1, 116) = 7.47, p = .007, \eta^2 = .06$. Females significantly increased their self-efficacy from pretest to posttest, whereas males did not show the increase. A contrast between Latinas and Caucasian males revealed a similar interaction pattern that revealed only the Latina's significant increase in their self-efficacy, $F(1, 63) = 4.98, p = .029, \eta^2 = .07$.

Mathematics learning in a socialized environment. The analysis of learning included 69 students, only those who had completed algebra posttests in both days. Table 2 presents the means and standard deviations of the pre- and posttests. The ANCOVA result revealed neither a main nor an interaction effect of student gender and ethnicity on learning, $F(1, 64) = 3.17, p = .080, \eta^2 = .05$. We also conducted a two-way repeated ANOVA to test the groups' learning gains over time. There was a significant main effect of the within-subject factor (time), $F(1, 65) = 53.28, p = .000, \eta^2 = .45$. A planned contrast between Latinas and Caucasian males did not reveal a significant difference in their learning gains. Overall, regardless of their gender and ethnicity, the student groups significantly improved their learning after working in the agent-based lessons.

To summarize, the ninth-grade females evaluated their agent significantly more positively than did males and the Latinos significantly more positively than did Caucasians. Second, the females showed significantly more positive attitudes toward the agent-based learning than did males. Third, the females significantly increased their mathematics self-efficacy after the agent-based learning, whereas the males did not show the increase. These gender differences in evaluations of an agent, attitudes, and self-

Table 1
Means and Standard Deviations for Posttest Learners' Evaluations of an Agent, Attitudes, and Self-Efficacy

Measure	Learner groups ($N = 120$)							
	Female				Male			
	Latina ($n = 30$)		Caucasian ($n = 26$)		Latino ($n = 29$)		Caucasian ($n = 35$)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Evaluation of their agent	86.3	4.13	69.23	4.43	77.17	4.20	54.57	3.82
Attitudes toward the lessons	9.04	0.50	7.93	0.55	7.56	0.51	7.19	0.45
Self-efficacy in the lessons	20.11	0.80	19.68	0.86	18.46	0.80	18.93	0.75

Table 2
Means and Standard Deviations for Pre- and Posttest Learning Measures

Measure	Learner groups (N = 69)							
	Female				Male			
	Latina (n = 14)		Caucasian (n = 18)		Latino (n = 14)		Caucasian (n = 23)	
	M	SD	M	SD	M	SD	M	SD
Learning								
Pretest	6.14	3.80	8.22	2.24	6.43	4.62	7.96	2.77
Posttest	9.29	4.12	10.78	1.86	7.64	5.20	11.04	2.65

efficacy were even more clearly manifested between the Latinas and the Caucasian males. Lastly, regardless of learner gender and ethnicity, all the groups significantly increased their learning after the lessons. Overall, the results suggested that the males and females have qualitatively different experiences in agent-based learning. In-depth interviews were conducted to better understand the nature of the students' experiences and the agent's characteristics that might most appeal to students of this age.

In-Depth Interview

Method

Interviewees. The interviews were focused on a deeper understanding of the females' experiences and the clear contrast between Latinas' and Caucasian males' reactions. Initially, 12 interviewees were randomly selected from the participant pool that had completed both lessons, which resulted in a sample of eight Caucasian males, two Caucasian females, and two Latinas. A second round of sampling was conducted to obtain a theoretical sample (six to eight) from each comparison group, to ensure a meaningful, thematic analysis. The sampling targeted the two female groups, from which six Caucasian females and four Latinas were further selected randomly.

Procedure. All interviews were conducted individually at the high schools and followed the lesson implementations. Three trained doctoral students conducted the interviews (each taking 20–30 min), using a loosely structured interview protocol that listed a set of main questions, for example, "What did you like or dislike about the lessons with a peer-like tutor?" and "What would you suggest for the improvement of the lessons?" The protocol allowed room for exploration and probing when necessary. To ensure the confidentiality of the interviewees, all identifiable comments were eliminated from the transcripts, and pseudonyms were used in the analysis.

Data analysis. The Constant Comparison method (Charmas, 2006) was used to identify major differences between the female and male groups. To ensure the quality and trustworthiness of the findings, the research team analyzed the interview data through a collaborative, reiterative process. Both authors read all transcripts individually and brainstormed several salient themes. Following that, the second author with the help of two graduate assistants launched a more systematic, thorough analysis, using the software Atlas ti (Version 6). Next, they summarized key information about the interviewees' experiences with the agent, their classroom experiences, and other critical information, such as the perceptions of

the agent, familiarity with computers, and the level of their attention to the agent. On the basis of recurring information in the summary, a list of open codes was developed. These codes were appended to relevant quotations in each transcript. A code output was generated. The team examined the output carefully to detect major patterns across the 22 interviewees and possible consistent relationships in the patterns. Lastly, the team elicited three main themes.

Results and Interpretation

Gendered perspectives and relationship building. The females and males demonstrated different views of their agent and developed different types of relationships with it while they engaged in the learning task. The males seemed to treat the agent as a mere tool and showed detached attitudes toward it, describing it as an "unnatural" or "fake" person and not being able to recall its name, gender, or ethnicity. They listed both positive and negative aspects of the agent. In most cases, their negative comments were longer and more varied than the positive ones. They found the agent's unsolicited explanations rather "annoying" and "boring." About half the males reported that they "turn[ed] off the voice," "skipped," or "ignored" narratives of their agent that they found not helpful. Mark's comments demonstrated this distant view of the agent:

Interviewer: So how did you find Chris [the agent] similar to a peer or friend?

Mark: I didn't really think of it as a friend, I just thought of it as like a little computer thing. (Interviewer: Oh really?) But yeah. I just don't really think that like computers are supposed to be your friend.

In contrast, the females seemed to treat their agent as if it were a friend or companion, and they built a humanlike, person-to-person relationship with their agent. They always called their agent by its name "Chris," used personal pronouns *she* or *he* to refer to the agent, and paid attention to various aspects of the agent (e.g., facial expressions, its hair style, the tone of speech). Not surprisingly, they reported their experiences with the agent-based learning very positively. This phenomenon was far more evident among the Latinas. Not one of the Latinas made negative comments about the agent; rather, all of them were effusive about their enjoyment in working with the agent. Perla (Latina) described her agent as being "really nice always" and "just like human thing . . . that someone is telling you compliments." Janet (Caucasian) said that her agent was "the person next to you [who] would help you with

whatever problem you need.” By and large, the development of a humanlike relationship with their agent seemed to generate positive effects on their learning process, but some negative consequences were also observed. Some girls were distressed by the agent’s negative feedback. The agent was “not friendly” but “mean” and “rude” and made them feel “hurt.” These girls seemed to project their interpersonal expectations onto the agent and were disappointed when their expectations were not met properly.

Consequences in learning: Students’ evaluations of agent effectiveness. Overall, both male and female students liked the agent’s immediate and individualized feedback. However, the males said that the explanations were sometimes redundant or, at other times, not specific enough. Although they valued the agent’s ability to provide feedback and/or to alert them to their mistakes, the males often skipped or turned off lengthy explanations to directly tackle the problem on their own. Only two males out of eight listed “good explanation” as a strength of the agent. The males’ complaints were mainly related to weak explanations not tailored “for me.” Rick’s comments exemplified the males’ reactions:

The only reason I marked that [evaluated negatively] I didn’t really like it ‘cause sometimes it explained like too much, like at the beginning of each section or something. It kept it kind of went on and on for me, so it just kind of got annoying for me to have to keep listening.

Conversely, the females spoke highly of the quality and relevance of the agent’s explanations. Almost all of them said that their agent had provided good explanations, which were “clear” and “very specific.” Selena said that the agent “explained every little part of it,” and “when I would get confused, she would explain what I did wrong clearly.” The girls rarely mentioned the actions often taken by the boys (e.g., skipping lengthy explanations). As a result, the girls were more likely to attend to and benefit from the coordinated instructional features (e.g., voice narration with the accompanying texts on the screen). Abby’s positive view consistently appeared in almost all females’ comments, “I would really enjoy it because like it explained it how to do it and it had visuals of how to do it, and it would explain how to go step by step. So it would be really helpful.” Presumably, the companionship that the females had built with the agent established a positive context for the subsequent learning process and made the females willing to listen to even lengthy explanations.

Connection between real and virtual contexts in learning experience. The gendered pattern of learning experiences with the agent did not seem to occur in a social vacuum. Rather, students’ views of their agent and the quality of their learning experiences with it seemed to be influenced by their everyday classroom experiences. The students who felt less supported in the classroom tended to develop positive attitudes toward the agent-based learning and reported positive learning experiences. The Caucasian males rarely expressed psychological stress in their classrooms; only two males showed a glimpse of social disconnection from their teacher. In contrast, all the females mentioned, at least once, a negative experience and/or feelings of insecurity in their classrooms. Most of them stated that their teacher did not care about their learning and was not willing to help them when they faced a difficulty.

This phenomenon was more manifested among Latinas. Whereas three Caucasian girls out of eight mentioned some positive aspect of the classroom, the Latinas’ narratives presented a greater disconnection in their relationships with the teacher and even a sense of fear and intimidation. Daniela (Latina) expressed discomfort with her teacher’s tone of voice: “They [teachers] teach you but sometimes you don’t get the thing and they teach again but in different voice.” In response to the question about the difference between the agent Chris and the teacher, Selena contrasted “upbeat and friendly” Chris with her “kind of intimidating” teacher. Although none of the Latinas showed difficulty with conversing in English, many indicated that the ordinary classroom instruction was “too fast,” and the teacher was “leaving you alone” even when students did not grasp the concepts. They felt relieved working with their agent, who would never blame them for not catching up to its speed. Some even argued, “You can learn more, and they [agents] teach you more, better than the teacher.”

To summarize, the interviews revealed that the males and females developed different relationship patterns with their agent. This gendered pattern of relationship building resulted in their differential evaluations of the quality and effectiveness of the agent’s feedback and explanations. Also, the students’ experiences with the agent were closely related to their everyday classroom experiences. The females, psychologically marginalized in the classroom, perceived the agent as a genuine companion who kindly helped them learn step by step.

General Discussion

This study was grounded in two theoretical premises. First, inequity issues in STEM education are attributable to the unsupportive context in STEM classrooms for traditionally underrepresented groups of students. To address this issue, educators should contrive supportive learning contexts, in which these students feel cared for and encouraged to engage in STEM learning. These contexts should accommodate the learning styles of the students who favor multifaceted interactions and social relations. Second, an embodied agent, with its social and empathetic capabilities, might afford humanlike interactions with the students. If designed carefully, agent-based learning could create a socially rich and inclusive context for those groups of students and, thereby, support their positive learning experiences and sustained intellectual pursuit in STEM. On the whole, the results from both phases of this study support the premises and show agent technology to be a promising tool in the resolution of urgent educational issues. The results also argue for the expansion of advanced learning technology.

Consistent with the present literature, the classroom experiment revealed clear gender differences in responses to agent-based learning. Females’ preference for social interactions and relationship building in the classroom seemed to be reflected consistently in their evaluations of their agent. Females rated the agent with an average of 4.6 on the 7-point scale, and males with an average of 3.9. In particular, Latinas rated the agent with an average of 5.1, and Caucasian males with an average of 3.2. The females’ favorable evaluations of their agent seemed to lead them to build more positive attitudes toward learning from the agent and to increase their self-efficacy after the lessons. Moreover, the females significantly increased their mathematics learning comparable to their

male counterparts after working with the agent. At a minimum, the conventional achievement gap favoring Caucasian males was not observed in the agent-based lessons. When the females realized that they had instructional support and were free from social embarrassment, they were more likely to engage and not be afraid of making mistakes, as indicated in the interviews.

The interviews supported the quantitative results of the classroom experiment and illuminated the nature of gender differences in the responses to the agent-based learning. First, both Latinas and Caucasian females engaged themselves in interactions with the agent and responded socially. Although the females admitted their agent to be a computer program, their interactions with it resembled their everyday social interactions with a friend in many ways. This implies that quality relationships must be important for females' mathematics learning even in technology-based environments. Regardless of a virtual or real space, relationship building is an essential and natural part of many females' learning process as explained by feminist scholars (Belenky et al., 1997; Noddings, 2003). The development of companionship with their agent provided the females with some advantages. It effectively engaged them in the task and let them be patient throughout the lessons. Also, it reduced the chance of experiencing the negative emotions that many females had in ordinary mathematics classrooms.

The study revealed that a persistent cultural and social disconnection existed between the females (more with Latinas) and their teachers. The students acknowledged that their teachers were overburdened with teaching a big class. Still, their feelings of disconnection were a challenge to their engagement and success in school mathematics (Lim, 2008). The features that they listed as supportive of their learning during the lessons were similar to the characteristics of *culturally relevant pedagogy* (Gay, 2000). The Latinas earnestly expressed their need for a psychologically "safe" space, where they could ask for help freely as many times as needed. The provision of a communal sense of learning—working together closely with someone willing to help—was a strength of the agent-based learning. Their feelings of connection to the agent and the agent's social encouragement seemed to lead them to full engagement in the task (Sciarra & Seirup, 2008).

The study also confirmed the trends in human–computer interaction and further extended our understanding in the area. The more computers present humanlike characteristics, the more likely they are to elicit social behavior from users (Lee, Jung, Kim, & Kim, 2006). Likewise, the agent Chris, looking peerlike, successfully elicited the females' social responses. Once the females identified their agent as a helper for their learning, it did not matter whether the helper was real or artificial (Turkle, 2011). More importantly, the study revealed that the boundary between real and virtual spaces was blurred. Students' online learning experience, either positive or negative, could be better understood in relation to their everyday classroom experience. Their learning experiences in the two spaces are closely interrelated, each providing an important context for the other and each influencing the other. In similar fashion, the females' (particularly the Latinas') positive experiences with the agent-based learning were influenced largely by their marginalized experiences in the everyday mathematics classrooms. An implication for the designers of advanced learning technology is that the careful observation and accurate understanding of challenges that students face in the classroom might be a

primary step in designing effective technology-based learning environments.

Several previous studies on embodied agents have reported that learners perceived a matched agent with their own gender or ethnicity more positively than a mismatched one (e.g., Kim & Wei, 2011; Moreno & Flowerday, 2006), indicating that social biases in the real world were consistently applied to agent–learner relations. However, our interest was in examining the potential of agent technology for countering existing stereotypes and biases. We focused on the motivational and persuasive role that a peerlike agent would play, regardless of its gender and ethnicity. Neither agent gender nor ethnicity was examined as a factor; instead, four versions of an agent differing in gender and ethnicity were randomly assigned to students, to control for a confounding effect by the learners' biased perceptions. In the interviews, the Latinas who worked with a Latino agent tended to express a higher level of affection for the agent than the Latinas who worked with a Caucasian agent. Nonetheless, all the Latinas agreed that their agent, either Caucasian or Latino, was a great helper.

Recently, there has been a growing awareness about the social and cultural aspect of females' and ethnic minorities' learning processes (Carr & Steele, 2009; Nasir et al., 2006). It is clear that more research is called for in designing effective learning technology for these students. This technology needs to support their identification with STEM topics and to include specific features that stimulate motivation. Technology-use trends in the United States show that African American and English-speaking Latino youths use Internet and mobile data more frequently than do Caucasian youths (Smith, 2010); thus, games and mobile technology could be a functional space for inviting these students' attention to STEM topics.

Lastly, the study had a few limitations. First, the agent in the study was designed to be a whole entity with instructional, social, affective, and aesthetic attributes. The interactions among these attributes and their relative contributions to the females' positive experiences should be clarified in the subsequent research. Second, both quantitative and qualitative data were collected in specific locations and with 2-day implementations. The findings should be generalized judiciously. Third, based on the results of the experiment, the interviews were focused on the contrast between the females and Caucasian males. Much is unknown in regard to Latino males' interactions with their agent and their experiences in the learning environment. Future research is warranted to overcome the limitations and confirm the findings of the present study.

References

- Arroyo, I., Murray, T., Woolf, B. P., & Beal, C. R. (2003). *Further results on gender and cognitive differences in help effectiveness*. Paper presented at the International Conference of Artificial Intelligence in Education, Sydney, Australia.
- Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, 94, 416–427. doi:10.1037/0022-0663.94.2.416
- Atlas ti (Version 6) [Computer software]. Retrieved from <http://www.atlasti.com/index.html>
- Bailenson, J. N., Yee, N., Blascovich, J., Beall, A. C., Lundblad, N., & Jin, M. (2008). The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. *Journal of the Learning Sciences*, 17, 102–141. doi:10.1080/10584400701793141

- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman.
- Bandura, A. (2006). Guide for constructing self-efficacy scales. In T. Urdan & F. Pajares (Eds.), *Self-efficacy beliefs of adolescents: A volume in adolescence and education* (pp. 307–337). Charlotte, NC: Information Age Publishing.
- Baylor, A. L. (2002). Expanding preservice teachers' metacognitive awareness of instructional planning through pedagogical agents. *Educational Technology Research & Development*, 50, 5–22. doi:10.1007/BF02504991
- Belenky, M. F., Clinchy, B. M., Golberger, N. R., & Tarule, J. M. (1997). *Women's way of knowing: The development of self, voice, and mind*. New York, NY: Basic Books.
- Bickmore, T. W. (2003). Relational agents: Effecting change through human-computer relationship (Unpublished doctoral dissertation). Massachusetts Institute of Technology, Boston, MA.
- Bloom, B. (1984). The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13, 4–16.
- Boaler, J. (2002). *Experiencing school mathematics: Teaching styles, sex, and setting*. Mahwah, NJ: Erlbaum.
- Carr, P. B., & Steele, C. M. (2009). Stereotype threat and inflexible perseverance in problem solving. *Journal of Experimental Social Psychology*, 45, 853–859. doi:10.1016/j.jesp.2009.03.003
- Charmas, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Thousand Oaks, CA: Sage.
- Cobb, P., & Yackel, E. (1998). A constructivist perspective on the culture of the classroom mathematics. In F. Seeger, J. Voigt, & U. Easchescio (Eds.), *The culture of the mathematics classroom* (pp. 158–190). London, England: Cambridge University Press.
- Crosnoe, R., Morrison, F., Burchinal, M., Pianta, R., Keating, D., Friedman, S. L., & Clarke-Stewart, K. A. (2010). Instruction, teacher–student relations, and math achievement trajectories in elementary school. *Journal of Educational Psychology*, 102, 407–417. doi:10.1037/a0017762
- Crosnoe, R., Rieggle-Crumb, C., Field, S., Frank, K. A., & Muller, C. (2008). Peer group contexts of girls' and boys' academic experiences. *Child Development*, 79, 139–155. doi:10.1111/j.1467-8624.2007.01116.x
- D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18, 45–80. doi:10.1007/s11257-007-9037-6
- du Boulay, B., Avramides, K., Luckin, R., Martínez-Mirón, E., Méndez, G. R., & Carr, A. (2010). Towards systems that care: A conceptual framework based on motivation, metacognition, and affect. *International Journal of Artificial Intelligence in Education*, 20, 197–229.
- Freeman, T. M., Anderman, L. H., & Jensen, J. M. (2007). Sense of belonging in college freshmen at the classroom and campus levels. *The Journal of Experimental Education*, 75, 203–220. doi:10.3200/JEXE.75.3.203-220
- Gay, G. (2000). *Culturally responsive teaching: Theory, research, and practice*. New York, NY: Teachers College Press.
- Geist, E. (2010). The anti-anxiety curriculum: Combating math anxiety in the classroom. *Journal of Instructional Psychology*, 37, 24–31.
- Graesser, A. C., Chipman, P., Haynes, B., & Olney, A. M. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48, 612–618.
- Gulz, A., & Haake, M. (2006). Design of animated pedagogical agents: A look at their look. *International Journal of Human-Computer Studies*, 64, 322–339. doi:10.1016/j.ijhcs.2005.08.006
- Iacobelli, F., & Cassell, J. (2007). Ethnic identity and engagement in embodied conversational agents. In J. G. Carbonell & J. Siekmann (Eds.), *Intelligent virtual agents* (pp. 57–63). Berlin, Germany: Springer. doi:10.1007/978-3-540-74997-4_6
- Jacobs, J. E., Davis-Kean, P., Bleeker, M., Eccles, J. S., & Malanchuk, O. (2005). "I can, but I don't want to": The impact of parents, interests, and activities on gender difference in math. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics: An integrative psychological approach* (pp. 246–263). New York, NY: Cambridge University Press.
- Kim, Y. (2007). Learners' expectations of the desirable characteristics of virtual learning companions. *International Journal of Artificial Intelligence in Education*, 17, 371–388.
- Kim, Y., & Wei, Q. (2011). The impact of user attributes and user choice in an agent-based environment. *Computers & Education*, 56, 505–514. doi:10.1016/j.compedu.2010.09.016
- Kinzie, M. B., & Joseph, D. R. D. (2008). Gender differences in game activity preferences of middle school children: Implications for education game design. *Educational Technology Research and Development*, 56, 643–663. doi:10.1007/s11423-007-9076-z
- Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30–43.
- Lee, K. M., Jung, Y., Kim, J., & Kim, S. R. (2006). Are physically embodied social agents better than disembodied social agents? The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies*, 64, 962–973. doi:10.1016/j.ijhcs.2006.05.002
- Lee, K. M., & Nass, C. (2003). *Designing social presence of social actors in human computer interaction*. Paper presented at the Computer Human Interaction (CHI), Ft. Lauderdale, FL.
- Lim, J. H. (2008). Double jeopardy: The compounding effects of class and race in school mathematics. *Equity & Excellence in Education*, 41, 81–97. doi:10.1080/10665680701793360
- Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, 136, 1123–1135. doi:10.1037/a0021276
- Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology*, 95, 419–425. doi:10.1037/0022-0663.95.2.419
- Moody, V. (2004). Sociocultural orientations and the mathematical success of African American students. *The Journal of Educational Research*, 97, 135–146. doi:10.3200/JOER.97.3.135-146
- Moreno, R., & Flowerday, T. (2006). Students' choice of animated pedagogical agents in science learning: A test of the similarity attraction hypothesis on gender and ethnicity. *Contemporary Educational Psychology*, 31, 186–207. doi:10.1016/j.cedpsych.2005.05.002
- Nasir, N. S., Rosebery, A. S., Warren, B., & Lee, C. D. (2006). Learning as a cultural process. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 489–504). Cambridge, England: Cambridge University Press.
- National Science Board. (2010). *Preparing the next generation of stem innovators: Identifying and developing our nation's human capital* (NSB-10-33). Washington, DC: Author.
- Neal, L., McCray, A. D., Webb-Johnson, G., & Bridgest, S. T. (2003). The effects of African American movement styles on teachers' perceptions and reactions. *The Journal of Special Education*, 37, 49–57. doi:10.1177/00224669030370010501
- Noddings, N. (2003). *Caring: A feminine approach to ethics and moral education*. Los Angeles: University of California Press.
- Petty, R. E., DeSteno, D., & Rucker, D. D. (2001). The role of affect in attitude change. In J. P. Forgas (Ed.), *Handbook of affect and social cognition* (pp. 212–233). Mahwah, NJ: Erlbaum.
- Ryokai, K., Vaucelle, C., & Cassell, J. (2003). Virtual peers as partners in storytelling and literacy learning. *Journal of Computer Assisted Learning*, 19, 195–208. doi:10.1046/j.0266-4909.2003.00020.x

- Sciarra, D., & Seirup, H. (2008). The multidimensionality of school engagement and math achievement among racial groups. *Professional School Counseling, 11*, 218–228. doi:10.5330/PSC.n.2010-11.218
- Smith, A. (2010). Mobile access 2010. *Pew Internet*. Retrieved from <http://www.pewinternet.org/Reports/2010/Mobile-Access-2010.aspx>
- Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology, 34*, 379–440. doi:10.1016/S0065-2601(02)80009-0
- Steffens, M. C., Jelenec, P., & Noack, P. (2010). On the leaky math pipeline: Comparing implicit math-gender stereotypes and math withdrawal in female and male children and adolescents. *Journal of Educational Psychology, 102*, 947–963. doi:10.1037/a0019920
- Stone, C. A. (1998). The metaphor of scaffolding: Its utility for the field of learning disabilities. *Journal of Learning Disabilities, 31*, 344–364. doi:10.1177/002221949803100404
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. New York, NY: Basic Books.
- Uekawa, K., Borman, K., & Lee, G. (2007). Student engagement in U.S. urban high school mathematics and science classroom: Findings on social organization, race, and ethnicity. *Urban Review, 39*, 1–43. doi:10.1007/s11256-006-0039-1
- Vanneman, A., Hamilton, L., Anderson, J. B., & Rahman, T. (2009). *Achievement gaps: How Black and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress*. Statistical analysis report (NCES 2009-455). Washington, DC: National Center for Education Statistics.
- White, B. Y., Shimoda, T. A., & Frederiksen, J. R. (1999). Enabling students to construct theories of collaborative inquiry and reflective learning: Computer support for metacognitive development. *International Journal of Artificial Intelligence in Education, 10*, 151–182.
- Willig, A. C., Harnisch, D. L., Hill, K. T., & Maehr, M. L. (1983). Sociocultural and educational correlates of success-failure attributions and evaluation anxiety in the school setting for Black, Hispanic, and Anglo children. *American Educational Research Journal, 20*, 385–410.

(Appendix follows)

Appendix

Agent Messages Examples

1. The agent presented persuasive (P) and informational (I) messages in the introductions to new sections and subsections.

At the start of the section on Distributing to Combine Like Terms, the agent said:

Hey, we are doing great. You know, if we do well in math, we can major in anything we want in college because many jobs require an understanding of math. Developing math skills now will give us

more opportunities later (P). Alright, in this section, we are learning how to distribute first and then combine like terms. Terms don't always come combined . . . (I).

2. The agent presented motivational and informational feedback in a sequence while students solved problems.

Question 5: Simplify the expression by distributing and combining the like terms.

$$4r + 4(r + s) = \underline{\hspace{2cm}}$$

Possible answers typed	Type of errors	Agent messages	
		Motivational feedback	Informational feedback
$8r + 4s$	None	<i>Excellent. Let's keep up the good work.</i>	—
$4r + 4r + 4s$	Like terms not combined	<i>Everybody makes a mistake. Let's learn from the mistakes we've made.</i>	<i>There is one more step after distributing. Let's check for like terms and combine them.</i>
$8r + s$	Partial distribution		<i>Distribute the 4 to both terms in the parentheses and then combine like terms.</i>
Any other (1st try)	Random		<i>First, distribute the 4 to both r and s in the parentheses, and then look for like terms.</i>
Any other (2nd try)	Random	<i>It was a challenge, but hang on there. It will pay off in the end.</i>	<i>Distribute the 4; that means we now have $4r$ plus $4r$ plus $4s$. The first two terms are alike, so we combine them and the answer is $8r$ plus $4s$.</i>

Received December 13, 2011

Revision received October 22, 2012

Accepted November 1, 2012 ■

Live Webcam Coaching to Help Early Elementary Classroom Teachers Provide Effective Literacy Instruction for Struggling Readers: The Targeted Reading Intervention

Lynne Vernon-Feagans and Kirsten Kainz
The University of North Carolina at Chapel Hill

Amy Hedrick
Lenoir Rhyne College

Marnie Ginsberg
The University of North Carolina at Chapel Hill

Steve Amendum
The North Carolina State University

This study evaluated whether the Targeted Reading Intervention (TRI), a classroom teacher professional development program delivered through webcam technology literacy coaching, could provide rural classroom teachers with the instructional skills to help struggling readers progress rapidly in early reading. Fifteen rural schools were randomly assigned to the experimental or control condition. Five struggling readers and 5 non-struggling readers were randomly selected from eligible children in each classroom. There were 75 classrooms and 631 children in the study. Teachers in experimental schools used the TRI in one-on-one sessions with 1 struggling reader in the regular classroom for 15 min a day until that struggler made rapid reading progress. Teachers then moved on to another struggling reader until all 5 struggling readers in the class received the TRI during the year. Biweekly webcam coaching sessions between the coach and teacher allowed the coach to see and hear the teacher as she instructed a struggling reader in a TRI session, and the teacher and child could see and hear the coach. In this way the classroom teacher was able to receive real-time feedback from the coach. Three-level hierarchical linear models suggested that struggling readers in the intervention schools significantly outperformed the struggling readers in the control schools, with effect sizes from .36 to .63 on 4 individualized achievement tests. Results suggested that struggling readers were gaining at the same rate as the non-struggling readers, but they were not catching up with their non-struggling peers.

Keywords: individualized instruction, literacy coaching, educational technology, rural classroom teacher, struggling readers

American schools have come under increasing scrutiny, largely because many children are not acquiring the skills they need to succeed in the larger culture (Grissmer, Flanagan, Kawata, & Williamson, 2000). The National Center for Education Statistics (2009) has reported that two thirds of fourth graders are not able to comprehend difficult texts, and 63% of fourth graders are reading

only at a very minimal level of proficiency. Of those families in poverty, only 28% of their children are reading at this minimum level of proficiency in fourth grade (Haager, Klingner, & Vaughn, 2007; Lyon, 2001). These low levels of reading proficiency are especially true for rural children from low-wealth communities who come to school with lower readiness skills than other children (Lee & Burkham, 2002). These lower readiness skills are due in part to the proportionately greater child poverty rates in rural versus urban areas with the gap between rural and urban poverty growing over the last 10 years (O'Hare, 2009). Since poverty is the most potent predictors of school success, even greater than mother education, two parent families, and a host of other demographic variables (Brooks-Gunn & Duncan, 1997), it is important to understand the context of schooling in these low wealth rural communities as well as develop and evaluate school programs that may be effective for children in the context of poverty.

The higher child poverty rate in rural communities impacts schooling, with a poorer tax base for schools, lower teacher pay, less educated teachers, and less access to educational resources (Amendum, Vernon-Feagans, & Ginsberg, 2011; Vernon-Feagans et al., 2012; Provasnik et al., 2007). When trying to improve student achievement, rural schools face challenges of geographic isolation and low population density that often lead to less ready access to state-of-the-art professional development for teachers

This article was published Online First September 9, 2013.

Lynne Vernon-Feagans, School of Education, The University of North Carolina at Chapel Hill; Kirsten Kainz, Frank Porter Graham Child Development Institute, The University of North Carolina at Chapel Hill; Amy Hedrick, Department of Psychology, Lenoir Rhyne College; Marnie Ginsberg, Frank Porter Graham Child Development Institute, The University of North Carolina at Chapel Hill; Steve Amendum, Elementary Education Department, The North Carolina State University.

Marnie Ginsberg is now in independent practice in Madison, Wisconsin.

Support for this research was provided by Institute of Education Sciences Grant R305A040056 for the National Research Center for Rural Education Support awarded to Tom Farmer and Lynne Vernon-Feagans.

Correspondence concerning this article should be addressed to Lynne Vernon-Feagans, School of Education, The University of North Carolina at Chapel Hill, 301K Peabody Hall, CB 3500, Chapel Hill, NC 27599-3500. E-mail: lynnevf@email.unc.edu

coupled with less access to technology in the classroom (Deweese, 2000; Vernon-Feagans, Gallagher, & Kainz, 2010). The issues faced by rural schools were underscored by a Government Accountability Office (2004) report that sampled rural school principals. The report highlighted rural school needs for better technology and teacher professional development.

The professional development program for classroom teachers evaluated in this article tries to address some of the needs of schools in rural low wealth schools with respect to both technology and teacher professional development. The Targeted Reading Intervention (TRI) provides teachers with professional development for struggling readers through state-of-the-art webcam coaching that allows literacy coaches thousands of miles away to provide real time feedback to teachers in their classrooms as the teachers instruct struggling readers. The program also provides extensive website materials for instruction, webcam workshops and webcam team/grade level meetings, as well as e-mail correspondence between teacher and coach.

Technology and Early Reading

Most of the previous research on the use of technology for early reading has focused on computer assisted instruction (CAI) developed for use by children who need or want additional instruction and practice in reading. This technology allows students to work on their own to supplement regular instruction in the classroom, minimizes teacher involvement, and has been shown to be effective in improving the early reading skills of children, including children with different skill levels and different ethnic and socioeconomic backgrounds (Blok, Oostdam, Otter, & Overmaat, 2002).

Recently, there has been particular emphasis on developing and examining CAI for children at risk for early reading disability (Chambers et al., 2011; Huffstetter, King, Onwuegbuzie, Schneider, & Powell-Smith, 2010; Saine, Lerkkanen, Ahonen, Tolvanen, & Lyytinen, 2011; Torgesen et al., 1999). These studies have demonstrated that children at risk for reading problems can progress in basic reading skills through CAI delivered by trained and specialized teachers/tutors in the resource room setting or in a mobile computer lab. One study demonstrated that an extended day program that used a web-based instructional framework was more effective than direct instruction delivered by a specialized teacher for children with significant reading delays in elementary school (Cole & Hilliard, 2006). Chambers et al. (2011) demonstrated that schools that used tutor-led small group instruction with a reading software could significantly improve the reading of struggling students in comparison to schools that did not use this tutor and software. Although these studies were important in underscoring the value of CAI for young at risk readers, they were likely costly if sustained because of the need for a specialized trainer or teacher who assisted the children during CAI.

Little research has focused on using technology to help the classroom teacher become more effective in instructing struggling readers except to introduce teachers to ancillary software that can supplement instruction in the classroom. A survey of elementary school teachers suggested that teachers used technology as a supplemental tool for instruction but did not use technology as the central tool for instruction (Franklin, 2007). Thus, studies of technology use by classroom teachers have assessed the effectiveness

of ancillary software packages for improving reading with mixed results as to the efficacy of such software for struggling readers. For instance, Lewandowski, Begeny, and Rogers (2006) found that at-risk elementary school readers practicing alone did not improve fluency, whereas both tutor- and computer-assisted groups of children significantly improved in reading speed and accuracy. Struggling readers who received training via computer performed as well as students who received individualized tutoring. On the other hand, Mathes, Torgesen, and Allor (2001) found that although the Peer Assisted Learning Strategies (PALS) reading program improved student reading, the addition of a phonological awareness computer software for struggling readers did not significantly improve reading over the traditional PALS program. Again, these computer programs probably saved time for the classroom teacher since the teachers did not have to be as involved in student learning but may have failed to help improve classroom teacher instructional literacy practices.

Some recent studies have focused on using technology to improve the teaching of preschool classroom teachers who are in Head Start or in pre-kindergarten programs for children from at risk backgrounds. These studies have used video and web based video platforms to promote effective professional development in literacy for these preschool teachers. In a series of studies examining the effectiveness of *My Teaching Partners*, teachers were asked to video themselves and then send the DVDs to a research team who in turn would give the teachers feedback on their classroom literacy practices in a few weeks or a month. Teachers also had access to a website for information on the program. This kind of professional development technology has proven effective for preschool teachers who serve a diverse group of learners (Mashburn, Downer, Hamre, Justice, & Pianta, 2010; Pianta, Mashburn, Downer, Hamre, & Justice, 2008). Especially interesting for the current study was Mashburn et al.'s (2010) study that compared two delivery systems to preschool teachers in a randomized control trial. Preschool teachers in the first condition had access to a literacy video library via a highly developed website with instructional materials for the teachers to easily access. The second condition allowed teachers access to the literacy video library but also allowed teachers to view their own teaching video clips on the website with reflective questions about their instruction. In addition, this group also participated occasionally in video conferencing with a literacy/language coach to discuss teaching practices. Mashburn et al. found that preschool classroom teachers who had both access to the video library but also had occasional coaching via the website and videoconferencing improved their children's vocabulary skills more than teachers who only had access to the video library. Another important recent study used a randomized control trial to examine the effectiveness of a literacy/language professional development program for preschool teachers called *Classroom Links to Early Literacy* under two different conditions: live literacy coaching of teachers versus video coaching (Powell, Diamond, Burchinal, & Koehler, 2010). In this case, both face to face and video coaching involved observing the teacher for 90 min every 2 weeks and giving her oral and written feedback in addition to written feedback on the videotaping of herself during teaching. This one semester study found that in both conditions, children who used their program had large gains in all areas of literacy and that there were no differences between the live and video conditions. This latter study suggests, as other

research without technology has found, that professional development with the addition of coaching may be the most effective way to improve the instruction of classroom teachers, especially in literacy.

A series of studies by Connor and colleagues (Al Otaiba et al., 2011; Connor et al., 2011; Connor, Morrison, & Petrella, 2004; Connor et al., 2009) have used technology in early elementary school in a different way to help classroom teachers. They have used software that individualizes instruction in reading for children and have used literacy coaches that help the classroom teacher use the software effectively in individualizing instruction. Their Instruction \times Skill interaction studies have been very innovative and have certainly shown that their software in conjunction with live literacy coaching in the classroom is very effective in helping all children progress in early reading.

These innovative technologies for delivering professional development to teachers in the form of a website that contained teacher videos of themselves teaching with later feedback and using innovative software to individualize instruction, although important, may be limited because classroom teachers were not able to receive immediate real-time feedback about their teaching practices with individual children that may be only accomplished through real time coaching of classroom teachers (Carlisle & Berebitsky, 2011; Elish-Piper & L'Allier, 2011; McGill-Franzen, Allington, Yokoi, & Brooks, 1999). In addition, these programs did not help teachers directly with their instruction of struggling students but rather focused on improving effective instruction for all children in the class.

Coaching, Early Reading, and the Targeted Reading Intervention

This recent research using technology supports the previous work on the importance of literacy coaching as a way to scaffold the skills of classroom teachers to make changes in classroom instruction. Research over the last 10 years has suggested that the most effective way to promote better teaching of reading by classroom teachers is by developing professional development programs that include the addition of ongoing support of teachers through the effective use of literacy coaches. Professional organizations like the International Reading Association (2004) and other research on coaching (Elish-Piper & L'Allier, 2011; McGill-Franzen et al., 1999; McKenna & Walpole, 2008) have demonstrated that materials and workshops alone are not enough to improve literacy instruction for classroom teachers, but the addition of having a literacy coach for the classroom teacher can improve teacher reading practices that are linked to improved student outcomes. For instance, Carlisle and Berebitsky (2011), in a quasi-experimental study of Reading First, found that professional development workshops alone were not as effective in improving first grade student decoding skills as professional workshops that also included literacy coaching over the school year. Coaches in this study visited classrooms and worked one-on-one with teachers to give feedback on their teaching, modeled methods of instruction, and served as a literacy resource. This real-time feedback for classroom teachers was available using previous technology (Mashburn et al., 2010; Pianta et al., 2008; Powell et al., 2010) because teachers' videos of themselves teaching requires extended time by literacy coaches/consultants to observe the vid-

eos and provide feedback to teachers. Furthermore, previous video coaching has had coaches watch the teacher for prolonged periods of time and then the teacher received delayed feedback on literacy instructional practices.

As Kennedy and Deshler (2010) have recommended, technology should be used only when it fits appropriately within the theory of change and enhances the underlying mechanisms of professional development in a positive way to maximize the possibility that teachers become more effective teachers for struggling readers. The delivery system of the Targeted Reading Intervention used a professional development program that targeted struggling readers. The TRI included the use of literacy coaching weekly through webcam technology that allowed the immediate feedback that is accomplished through live one-on-one literacy coaching. That is, coaches thousands of miles away were able to see and hear the classroom teacher as she provided reading instruction to a struggling reader and the coach could give the teacher real time feedback on practices as well as problem solve about the best strategies to use with a particular struggling reader. This webcam approach to help classroom teachers instruct their struggling readers could help avoid the need for a specialized teacher to implement remedial reading programs (Amendum et al., 2011). Using webcam technology may also be more cost effective and feasible in rural areas where geographic isolation may prevent access to high quality professional development (Vernon-Feagans et al., 2012; Provasnik et al., 2007).

Thus, the intervention described in this study (The Targeted Reading Intervention) was developed in order to provide classroom teachers with particularly effective reading strategies for struggling readers in early elementary school. These strategies were implemented with the help of a literacy coach who worked with the teacher so she learned the strategies in instructional one-on-one diagnostic teaching sessions so the teacher could see the progress of individual struggling readers. Technology was used that allowed the literacy coaches to see and hear the teachers in these one-on-one sessions and give real time feedback to maximize teacher instructional change. Within our more elaborated model of teacher change, we included coaches who scaffolded the experience of teachers as the teacher worked individually with one struggling reader in the hope of changing the way the teacher delivered instruction to struggling readers. Thus, teacher experience of being coached and working with one child at a time has been hypothesized to be one mechanism for improving effective teacher instruction (Morgan, Timmons, & Shaheen, 2006; Risko et al., 2008).

Summary

The TRI has a number of unique elements that together may create the most effective instruction for struggling readers within the regular classroom setting. First, unlike many other interventions, the TRI uses the classroom teacher to deliver the intervention to each individual struggling reader through efficient, diagnostic one-on-one instructional sessions. Second, the TRI iterative process of the teacher working with one struggling reader at a time helps the teacher understand and experience the success as she sees the struggling reader make rapid gains. Third, the TRI uses an innovative, web-based, collaborative coaching model. Biweekly,

each TRI teacher uses a laptop computer with a webcam in her classroom so that she can see and hear her literacy coach and the coach can see and hear her working with an individual struggling reader. Real time feedback and problem solving can be employed during these live sessions for individual children.

In this study, we sought to examine whether the TRI could accelerate struggling readers' early literacy skills so that they not only made significant progress across a year but that they began to catch-up to their non-struggling classroom peers. This required examination of struggling and non-struggling readers' performance in the intervention schools relative to each other and to struggling and non-struggling readers in the control schools. The research questions were the following: (1) Do struggling readers who participate in TRI demonstrate better performance on tests of early literacy at the end of a school year than struggling readers who do not participate in TRI?; (2) When compared to struggling readers in control schools and to non-struggling classroom peers, does the spring performance of struggling readers in the intervention schools indicate that they are catching up to their non-struggling classroom peers?

Method

Setting

Sixteen rural schools from five poor rural counties in different regions of the United States participated in the study, including schools in Texas, New Mexico, Nebraska, and North Carolina. All kindergarten and first grade classrooms in each school participated. Schools within each school district were pair matched on the following: percentage of free and reduced lunch, school size, percentage of minorities, and participation in *Reading First*. One member of each pair was randomly selected to be the experimental school. Difficulties with accessing the Internet led to the withdrawal of one small experimental school that contained one kindergarten and one first grade classroom. The 15 remaining participating schools included 75 kindergarten and first grade classrooms and 631 students. All schools received Title I funding.

Participants

The demographics of the 631 children who participated in the fall assessments are described in Table 1. Since all schools were in low-wealth counties, the reported maternal education of these children was generally just beyond high school. Approximately 50% of the children were from minority backgrounds, and half were boys. Teacher demographics are shown in Table 2 and are consistent with literature on rural schools. Teachers had more years of experience than reported for urban teachers, with an average of 15 years of teaching experience (Lee & Burkham, 2002).

Within each experimental and control classroom, teachers identified children who were struggling and non-struggling readers with the help of the TRI literacy coach, mandated state assessment data and classroom performance within 2 months of the beginning of the school year. Teachers then rated all the children in the class as to whether they were profiting from regular classroom instruction in reading and were on grade level. Based on this information, five struggling readers were randomly selected from those children

rated as significantly below grade level in reading, and five non-struggling children were randomly selected from those children who were rated as on or above grade level in each classroom. Because of a variety of permission and attrition factors, there were approximately nine children who participated in each classroom.

We defined four groups of children for analysis purposes and to test hypotheses about the effectiveness of the TRI: struggling readers in experimental intervention schools (SRI), non-struggling readers in the same experimental intervention schools (NSI), struggling readers in non-intervention control schools (SRC), and non-struggling students in the same non-intervention control schools (NRC). In experimental intervention schools, there were 192 struggling readers (SRI) and 203 non-struggling readers (NSI). In the control schools, there were 107 struggling readers (SRC) and 129 non-struggling readers (NRC).

The Targeted Reading Intervention Using Webcam Coaching

The main objective of the overall Targeted Reading Intervention (TRI) was to help the classroom teacher acquire key reading diagnostic strategies to promote rapid reading gains in K-1 struggling readers through a technology driven professional development program that included ongoing biweekly coaching from a literacy consultant. The coaches all had extensive experience as teachers and/or reading coaches in early elementary school. Most were doctoral students in the School of Education. These coaches went through an intensive training that included videotaping themselves working with individual children and receiving feedback from the intervention director of the project. Finally, coaches were given feedback throughout the academic year with respect to the challenges of working with teachers who were not always motivated to implement the TRI, coincident with current literature on coaching (Al Otaiba, Hosp, Smartt, & Dole, 2008).

All teachers in the experimental group received a 3-day summer workshop to learn the TRI strategies and to practice them. The intervention director and the trained reading coaches led the 3-day institute. During the year, a literacy coach used cost effective webcam technology to meet with the teacher for about 20 min every 2 weeks over the instruction of an individual struggling reader. When the student made rapid progress, the student was transitioned to a small group, and another child was chosen to work one-on-one with the teacher. Through this webcam technology, the literacy coaches could help the classroom teacher use the TRI strategies effectively with each struggling reader in real time, help decide when a student was ready to be transferred to a small group session, and problem solve about students who were not making rapid progress. In addition, the literacy coach also met with each school team for 30 min bi-weekly through webcam technology to further reinforce the strategies and problem solve about individual children. Finally, workshops were also provided to the teachers every few months via webcam to support their developing understanding of the TRI process, models, and strategies. The TRI protected website contained all the training videos, instructions, and manuals that could be downloaded by teachers and links to downloadable books and so forth.

During the school year, the teachers implemented the TRI in 15-min one-on-one sessions with a struggling reader that included the following three parts each day:

Table 1
Child Demographics and Achievement Scores

Variable	Statistic	Kindergarten				First grade			
		NSC	SRC	NSI	SRI	NSC	SRC	NSI	SRI
Male	<i>N</i>	59	57	94	90	70	50	109	102
	<i>%</i>	0.59	0.63	0.55	0.54	0.50	0.64	0.33	0.54
White	<i>N</i>	59	57	94	90	70	50	109	102
	<i>%</i>	0.59	0.39	0.52	0.54	0.57	0.40	0.47	0.45
Maternal education	<i>N</i>	56	53	91	81	67	47	101	95
	<i>M</i>	14.18	12.64	13.63	13.06	13.67	13.15	13.52	12.99
	<i>SD</i>	1.99	2.25	2.13	2.33	2.16	2.03	2.40	2.37
Fall PPVT-III	<i>N</i>	53	54	94	90	67	48	109	102
	<i>M</i>	102.81	94.09	98.67	91.50	98.79	88.54	98.14	91.18
	<i>SD</i>	13.44	14.91	12.92	14.98	14.71	15.43	13.16	12.76
Fall WA score	<i>N</i>	58	57	94	89	69	50	107	102
	<i>M</i>	426.41	411.12	431.16	409.67	468.06	450.96	470.28	455.02
	<i>SD</i>	18.66	18.87	22.01	23.11	19.21	18.99	17.03	19.76
Fall LW score	<i>N</i>	59	57	94	90	68	50	109	102
	<i>M</i>	372.19	353.88	376.90	354.60	424.90	401.30	431.06	406.10
	<i>SD</i>	20.95	20.36	22.45	22.45	27.24	19.02	25.63	19.51
Fall PC score	<i>N</i>	59	57	94	90	69	50	109	102
	<i>M</i>	411.88	403.54	409.29	402.66	449.48	431.24	454.93	431.38
	<i>SD</i>	17.51	13.31	18.98	13.60	26.54	17.34	20.85	20.19
Fall SS score	<i>N</i>	59	57	89	88	70	50	109	102
	<i>M</i>	464.76	445.61	468.22	448.78	490.50	483.50	491.36	484.65
	<i>SD</i>	18.72	18.67	14.87	17.61	8.32	11.81	7.54	9.12
Spring PPVT	<i>N</i>	55	51	88	84	69	45	103	92
	<i>M</i>	105.15	96.63	100.50	95.62	103.29	94.38	100.51	91.67
	<i>SD</i>	17.12	12.68	13.85	11.50	15.19	15.42	15.44	14.61
Spring WA score	<i>N</i>	55	51	88	84	69	45	103	92
	<i>M</i>	460.38	449.10	465.24	456.81	482.64	466.73	484.93	474.21
	<i>SD</i>	17.18	20.83	17.95	21.41	15.35	17.02	19.05	16.77
Spring LW score	<i>N</i>	55	51	88	84	68	42	103	92
	<i>M</i>	408.78	389.71	418.39	403.04	458.49	434.40	463.65	442.90
	<i>SD</i>	21.09	16.21	22.70	22.38	23.71	20.63	22.05	19.43
Spring PC score	<i>N</i>	55	51	87	84	69	45	103	92
	<i>M</i>	435.96	417.04	445.06	428.60	472.84	456.07	475.01	461.92
	<i>SD</i>	22.73	18.64	22.09	21.38	13.09	17.21	12.43	14.45
Spring SS score	<i>N</i>	55	51	88	84	69	45	103	92
	<i>M</i>	484.44	477.63	489.73	483.93	498.75	491.93	497.18	494.36
	<i>SD</i>	10.41	13.83	6.87	10.70	5.11	8.92	8.45	6.24

Note. NSC = non-struggling readers in control schools; SRC = struggling readers in control schools; NSI = non-struggling students in intervention schools; SRI = struggling readers in intervention schools; PPVT-III = Peabody Picture Vocabulary Test—III; WA = Word Attack; LW = Letter Word Identification; PC = Passage Comprehension; SS = Spelling of Sounds.

1. Re-reading for fluency. The teacher asks the student to re-read a selection that she/he has read at least once in the recent past for the purpose of developing reading fluency. The teacher might model fluent reading with some of the text, depending on the skill level of the child. This is done even with children who are non-readers through scaffolding and modeling. For example, asking the child where to start reading and identifying initial sounds in words can be a way to help a beginner be successful, even when they have extremely limited alphabetic knowledge.

2. Word work. This innovative approach provides the teacher with a variety of assessment-based multi-sensory instructional strategies for helping the child manipulate, say, and write words. In the early stages, there are four major strategies that are employed using a white board and letter sounds (letter combinations) tiles to help children make words and to see, hear, and manipulate differences between words. These four strategies were adapted to four major levels of child skill in reading and writing words. Level 1 of Word Work was geared to children who had almost no knowledge

Table 2
Demographics of the Teachers

Variable	Treatment (<i>n</i> = 43)			Control (<i>n</i> = 32)		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Race						
Black/African American	6			4		
White/European American	35			25		
Other	1			3		
Missing	1					
Gender						
Female	41 ^a			32		
Age						
20–29	6			8		
30–39	11			8		
40–49	12			8		
50–59	11			7		
60+	3			1		
Certification level						
Elementary education certified	40			28		
Master’s degree or higher	10			22		
Experience						
Total years teaching	17.60	10.77		13.33	9.69	
Total years teaching current grade	8.95	7.88		5.31	8.95	
Total years teaching at current school	8.00	5.59		7.45	8.00	
Total years teaching in current county	12.53	8.93		9.45	8.56	

^a One teacher not reporting gender, and one male teacher in experimental schools.

of the alphabetic principle and focused on three-sound words with short vowels. Level 2 of Word Work was geared to slightly more advanced knowledge of the alphabetic principle and introduced children to four-sound words. The third level of Word Work allowed children more advanced phonics work with long vowel sounds that can be represented by a variety of vowel constellations. The fourth and final level of Word Work focused on multi-syllabic words.

Along with the help of their literacy coaches, teachers made decisions about when to progress to more challenging levels of word identification and adopt slightly different strategies. The graphic organizer for the teacher helped her understand the four levels and the key diagnostic criteria to place a child within these skill levels. Thus, teachers learned to assess the child’s level of word identification and select a particular diagnostic strategy that matched the skill level of the child to achieve *instructional match* (Bear, Invernizzi, Templeton, & Johnston, 2003; Beck, 2006; Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007; Connor et al., 2009; Morris, Tyner, & Perney, 2000). All TRI strategies demonstrate the alphabetic principle, help students learn phoneme–grapheme (sound–symbol) relationships, develop students’ segmenting and blending abilities (phonemic awareness tasks), and help students recognize sight words. The four primary strategies are (1) *Segmenting Words*; (2) *Change One Sound*; (3) *Read, Write, and Say*; and (4) *Pocket Phrases*.

Segmenting Words helps children to acquire knowledge about the sounds in simple, but progressively more difficult words, by allowing the child to use the letter sound tiles to build words by saying and moving each tile. For example, a child with limited alphabetic knowledge would begin at the lowest level, targeting

words that can be made with the following few letters: a, s, m, t, and p. Three-letter words with beginning consonants, such as “sat” or “mop” and not “top,” would be chosen because these types of words are the easiest to teach phoneme segmentation since the teacher can stretch out the sounds more easily. For example, the teacher might place three letter-sound cards (t, m, a) at the top of the Word Work board and say,

Hannah, I want you to help build a word right here [tapping lines on bottom of board]. The first word is “mat,” /mmmat/ [as she drags her finger along the three short lines at the bottom of the board in concert with the sounds she is making]. I wiped my feet on the *mat*. What sound do you hear *here* [pointing to Word Work board] in the word /mmmat/?

With feedback, the child will progress and this allows the teacher to gradually progress to more challenging words.

Change One Sound helps children contrast the sounds between words by placing selected letter sound tiles in front of the child and helping the child to make a word like “map” from the letter sound tiles and then asking him to change that word to “mop” by replacing the “a” tile with the “o” tile. As children become more proficient in this strategy, teachers may focus on medial, beginning or ending sounds and always in the contexts of real words.

Read, Write, and Say helps children to read new words and write those words on the white board. As children write the words they also say the words again.

Pocket Phrases helps children to remember sight words/phrases by writing these words/phrases on cards and asking them to show and read to others in their class or at home.

These four strategies, along with other more advanced strategies, can be used with TRI instructional levels that gradually expose the student to more and more alphabetic complexity, keeping him/her challenged. After each session, the teacher then goes back to her diagnostic map and develops a plan for the child’s next session.

3. Guided Oral Reading (GOR). Strategies are employed in a text chosen at the child’s instructional reading level, as guided by the *Word Work* sessions and *Diagnostic Map*. Teachers pay particular attention to scaffolding children’s abilities to summarize, predict, make connections, and inferences. Vocabulary words that may be difficult are defined and a picture dictionary is available during this part each session. During a book-reading session the teacher might ask for the child to define a word, to answer what might happen next, or to answer a causal question about the storyline. Having children orally summarize the story at the end helps the teacher understand if the child truly understood the book as well as whether the child understands the conventions of storytelling. Having teachers ask concrete and abstract questions about the story also can help them understand whether the child understands the nuances of the story and help them understand whether the child understands what is demanded by different levels of questions.

Data Collection and Measures

All children in the study were administered a battery of standardized tests in the fall and again in the spring of the school year. Teachers filled out questionnaires about their professional background and classroom. All child assessments were done in the

schools in a quiet room. Trained graduate students or former teachers conducted the child assessments. The assessors participated in a 2-day training, which included the administration of the complete battery with non-participating students. Assessors were not informed which schools were experimental or control. The following measures were administered to children in the fall and the spring.

Four subtests of the *Woodcock–Johnson III Diagnostic Reading Battery* (WJ-DRB III; Woodcock, Mather, & Schrank, 2004) were administered to all children. *Word Attack* measures skill in applying phonic and structural analysis skills to the pronunciation of unfamiliar printed sounds and words. The initial items require the child to produce sounds for single letters. The remaining items require the child to read aloud letter combinations that are phonetically consistent, or regular, patterns in English orthography but are non-words or low-frequency words. The items become progressively more difficult. *Word Attack* has a median reliability of .87 in the 5–19 age range (Woodcock et al., 2004).

Letter Word Identification measures the child's word identification skills. The initial items require the child to identify letters that appear in large type, and the remaining items require the child to pronounce words correctly. The items become increasingly difficult as the selected words appear less and less frequently in written English. *Letter Word Identification* has a median reliability of .91 in the five-to-19 age range (Woodcock et al., 2004).

Passage Comprehension initial items measure symbolic learning and require the child to match a rebus with an actual picture of an item. The more advanced items employ a modified cloze procedure that requires the child to read a short passage and provide a missing key word which makes sense within the context of the passage. The items become increasingly difficult by removing pictorial support and by increasing passage length and difficulty as well as vocabulary complexity. *Passage Comprehension* has a median reliability of .83 (Woodcock et al., 2004).

Spelling of Sounds measures the child's spelling ability, in particular, phonological and orthographical coding skills. Initial items require the child to write single letters for sounds. Remaining items require the child to spell letter combinations that are regular patterns in English. Items increase in difficulty by requiring more complex spelling patterns. *Spelling of Sounds* has a median reliability of .74 (Woodcock et al., 2004).

The *Peabody Picture Vocabulary Test—III* (PPVT—III; Dunn & Dunn, 1997) is an individually administered, norm-referenced test of receptive vocabulary knowledge. Children are asked to select a picture that best represents the meaning of the stimulus word presented orally by the examiner. Alpha coefficients for the PPVT—III for elementary age students range from .92 to .95 (Dunn & Dunn, 1997).

Fidelity of Implementation

To assess fidelity of implementation of the TRI, classroom teachers reported exposure of each target child to the TRI as well as the teachers' adherence to the elements of the TRI to an on-site facilitator during the biweekly team meetings with the literacy coach. The teachers then entered the data online. The Fidelity data are summarized in Table 3 across kindergarten and first grade since there were no grade level differences on any of the fidelity measures.

Exposure was measured by the number of weeks that each child received the TRI over the course of the year and the total number of sessions per week. Each of the five target children received one-on-one TRI intervention an average of 6 weeks with 2.4 sessions per week for a total of about 14 sessions per child.

Adherence to the TRI was measured by the number of reported weeks that each of the three parts of the TRI were implemented with each child: *Re-Reading for Fluency*, *Word Work*, and *Guided Oral Reading*. Teachers reported that 80% of the week's *Re-Reading for Fluency* was implemented, 96% of the week's *Word Work* was implemented, and 92% of the week's *Guided Oral Reading* was implemented.

Results

Missing Data Methods

More than 85% of the sample participated in fall and spring assessments and provided demographic background information. To avoid imprecise regression estimation due to missing data, we created and analyzed multiple imputed data sets in SAS Version 9.1. Multiple imputation procedures use an iterative (chained equations) method to estimate the multivariate relations among study variables for cases with available data. These observed relations among study variables are then used to estimate plausible values for missing data. Creating multiple data sets with plausible values for missing data and aggregating solutions from analyses using multiple data sets provides the best approximation of relations among variables given no missing data (Graham, Olchowski, & Gilreath, 2007; Schafer & Graham, 2002). Consequently, the analysis of variance (ANOVA) and analysis of covariance (ANCOVA) models presented below were run on each of 20 imputed data sets, and model parameters were aggregated across the data sets using the PROC MIANALYZE function in SAS. The imputation model included the following: fall and spring assessment scores for all outcomes, child grade, child race (White, Black), child gender, mother's education, and dummy variables indicating school identification and randomized treatment status.

Preliminary Analysis

Before testing intervention effects on student literacy outcomes, we conducted preliminary analysis to verify the validity of teachers' identification of struggling readers. To validate teachers' identification, we compared fall scores on all outcomes of interest for struggling readers and non-struggling readers in the sample. These models were estimated in SAS 9.1 as three-level ANOVAs

Table 3
Fidelity of Implementation

Variable	N	M	SD
Total number of weeks of TRI	167	6.02	3.77
Number of sessions per week of the TRI	167	2.39	0.79
Proportion of weeks Re-Reading for Fluency done	167	0.83	0.25
Proportion of weeks Guided Oral Reading done	167	0.86	0.23
Proportion of weeks Word Work done	167	0.97	0.09

Note. Teachers with intermittent reporting of fidelity were dropped from the fidelity analysis. TRI = Targeted Reading Intervention.

accounting for the nesting of students with classrooms and classrooms within schools. Three-level models predicted fall scores as a function of a four-category intervention group variable at Level 2 (SRC, NSC, SRI, NSI). Follow-up contrasts of struggling versus non-struggling fall scores were conducted to test for mean differences in performance before intervention. For all outcomes of interest, struggling readers scored significantly lower than non-struggling readers did before this intervention. We present the results of the tests of mean differences in Table 4.

Tests of the Effects of the Intervention

Analytic strategy. Multi-level (hierarchical) models were used to examine our questions about the effectiveness of the TRI for struggling readers. Separate models were conducted for each of five outcomes: Word Attack, Letter Word Identification, Passage Comprehension, Spelling of Sounds, and PPVT. All models were estimated in SAS Version 9.1 as a three-level ANCOVA accounting for the nesting of students within classrooms and classrooms within schools. Effect sizes for significant treatment effects were calculated by dividing the contrast coefficient (mean difference) by the square root of total variation in the model.

The three-level ANCOVA predicted spring scores as a function of fall pre-test scores as a fixed effect at Level 1, a four-category intervention group fixed effect at Level 2, and a set of level-one fixed effects used as covariates across all models: gender is male, mother's years of education, grade ($K = 0$, first = 1), and race is White. This model estimated random effects for classroom and school intercepts. All covariates including the pre-test were centered for analysis, so that the intercept in the models reflected average spring scores for the treatment reference group, that is, struggling readers in intervention schools.

Intervention effects to answer Question 1 were established by testing the significance of the conditional mean difference between spring scores for struggling readers in intervention schools and struggling readers in control schools.

In order to answer Question 2 about whether our experimental intervention children were progressing at the same rate as their non-struggling peers, we used the following rationale. Considering that the performance of students in control schools represented expected performance for students in experimental schools if intervention were not administered, we proposed that evidence for catch-up would be clearly established given four effects: (1) significant and positive intervention effects between the SRI and the

SRC; (2) non-significant differences between conditional spring scores for struggling readers in intervention schools (SRI) and non-struggling readers in intervention schools (NSI; i.e., controlling for fall scores students in intervention schools are changing at a similar rate regardless of struggling status); (3) significant differences between conditional spring scores for struggling readers in control schools (SRC) and non-struggling readers in control schools (NSC; i.e., non-struggling readers are changing at a greater rate than struggling readers are in control schools); and (4) non-significant differences between conditional spring scores for non-struggling students in intervention schools (NSI) and non-struggling readers in control schools (NSC; i.e., no evidence that non-struggling students in intervention schools are lagging compared to non-struggling students in control schools). These contrasts are presented in Table 5.

The reduced form equation for the model is as follows:

$$Y_{ijk} = \gamma_{000} + \gamma_{100}(\text{pre-test})_{ijk} + \gamma_{020}(\text{treatment})_{jk} + \gamma_{300}(\text{male})_{ijk} \\ + \gamma_{400}(\text{mother's education})_{ijk} + \gamma_{500}(\text{grade})_{ijk} + \gamma_{600}(\text{White})_{ijk} \\ + u_{00k} + r_{0jk} + e_{ijk}$$

In this notation, fixed effects are represented by gammas (γ), and random effects are reflected in two error terms: a term for Level 3 variation between schools (u_{00k}) and a term for Level 2 variation between classrooms in schools (r_{0jk}). In exploratory models, we tested a Grade \times Treatment interaction. Because that interaction was not significant for any of the outcomes, we excluded it from the final models. Results from the multi-level ANCOVA appear in Table 5. Formal tests of treatment main effects are represented twice in the table: first as the coefficient for the struggling control group listed in the fixed effects (SRC) and second as the formal contrast of struggling intervention (SRI) and struggling control (SRC) conditional means.

The table contains fixed effects, variance components, and group contrasts obtained through estimate statements for each of five outcomes. Within Table 5, we provide variance components for each model. Significance tests of the variance components indicated that the variation between schools was not significantly different from zero for any of the five outcomes. The variation between classrooms within schools was significantly different from zero for Letter Word Identification (LW), Passage Comprehension (PC), and PPVT-III only.

Word Attack. Controlling for differences in pre-test scores, TRI had a positive effect on struggling readers' Word Attack skills. Spring scores for struggling readers in intervention schools were 5.65 higher than scores for struggling readers in control schools ($b = 5.65$, $p = .04$). There was some evidence that TRI promoted catch-up for Word Attack skills. Non-struggling readers in intervention schools did not outperform struggling readers in intervention schools controlling for fall performance ($b = -0.55$, $p = .75$). However, non-struggling readers in control schools did outperform struggling readers in those schools, controlling for fall scores ($b = -5.13$, $p = .02$). There was no evidence that non-struggling readers in intervention schools underperformed relative to non-struggling readers in control schools ($b = 1.08$, $p = .64$). Above and beyond intervention effects, male students had lower spring Word Attack scores, and more maternal education was associated with higher spring scores.

Table 4
Pretest Differences Between Struggling and Non-Struggling Readers

Variable	NS versus SR	SE	p
Fall WA	16.79	1.56	<.0001
Fall LW	21.13	1.67	<.0001
Fall PC	13.37	1.49	<.0001
Fall SS	12.73	1.06	<.0001
Fall PPVT-III	7.43	0.98	<.0001

Note. NS = non-struggling readers; SR = struggling readers; WA = Word Attack; LW = Letter Word Identification; PC = Passage Comprehension; SS = Spelling of Sounds; PPVT-III = Peabody Picture Vocabulary Test-III.

Table 5
HLM Intervention Effects and Planned Comparisons

Fixed effects	WA				LW				PC			
	B	SE	p	d	B	SE	p	d	B	SE	p	d
Pretest	0.45	0.03	<.0001		0.64	0.03	<.0001		0.32	0.03	<.0001	
White	0.55	1.45	.71		0.66	1.42	.64		3.33	1.47	.02	
Male	-4.32	1.27	.00		-3.64	1.25	.00		-2.60	1.32	.05	
Maternal education	0.73	0.31	.02		0.64	0.31	.04		1.45	0.32	<.0001	
Grade	0.81	2.11	.70		11.43	2.24	<.0001		22.92	2.06	<.0001	
NSC	-0.53	2.68	.84		-0.03	3.11	.99		3.65	2.81	.20	
SRC	-5.66	2.72	.04		-8.39	3.20	.01		-7.65	2.88	.01	
NSI	0.55	1.68	.74		2.38	1.64	.15		8.29	1.63	<.0001	
Variance components												
Level 3 variation	10.44	8.81	.24		19.54	12.07	.11		13.31	10.59	.21	
Level 2 variation	16.73	9.01	.06		21.42	9.73	.03		18.21	8.55	.03	
Level 1 variation	217.69	13.89	<.0001		198.31	12.93	<.0001		216.55	13.76	<.0001	
Contrasts												
SRI vs. SRC	5.66	2.72	.04	0.36	8.39	3.20	.01	0.54	7.65	2.88	.01	0.48
SRI vs. NSI	-0.55	1.68	.75	0.04	-2.38	1.64	.15	0.15	-8.29	1.63	<.0001	0.53
SRC vs. NSC	-5.13	2.13	.02	0.33	-8.36	2.09	<.0001	0.54	-11.30	2.11	<.0001	0.72
NSI vs. NSC	1.08	2.62	.64	0.07	2.41	3.05	.43	0.16	4.65	2.77	.09	0.29

Fixed effects	SS				PPVT-III			
	B	SE	p	d	B	SE	p	d
Pretest	0.31	0.02	<.0001		0.65	0.04	<.0001	
White	0.78	0.71	.27		2.92	1.00	.00	
Male	-1.20	0.62	.05		0.78	0.84	.35	
Maternal education	0.50	0.15	.00		0.41	0.22	.07	
Grade	2.06	1.03	.05		-0.37	1.36	.79	
NSC	-0.98	1.30	.45		4.01	2.03	.05	
SRC	-3.23	1.33	.02		1.34	2.12	.53	
NSI	-0.12	0.83	.89		2.67	1.01	.01	
Variance components								
Level 3 variation	2.63	1.99	.19		5.57	5.21	.29	
Level 2 variation	2.78	1.79	.12		22.85	6.49	.00	
Level 1 variation	51.01	3.25	<.0001		86.48	5.57	<.0001	
Contrasts								
SRI vs. SRC	3.23	1.33	.02	0.63	-1.59	2.13	.46	0.15
SRI vs. NSI	0.12	0.83	.89	0.02	-2.38	0.98	.02	0.22
SRC vs. NSC	-2.25	1.05	.03	0.44	-2.60	1.30	.05	0.24
NSI vs. NSC	0.87	1.26	.49	0.17	-1.80	2.08	.39	0.17

Note. Bolded *ds* are significant effect sizes. HLM = hierarchical linear modeling; WA = Word Attack; LW = Letter Word Identification; PC = Passage Comprehension; NSC = non-struggling readers in control schools; SRC = struggling readers in control schools; NSI = non-struggling students in intervention schools; SRI = struggling readers in intervention schools; SS = Spelling of Sounds; PPVT-III = Peabody Picture Vocabulary Test—III.

Letter Word Identification. Controlling for differences in pre-test scores, TRI had a positive effect on struggling readers' Letter Word Identification skills. Spring scores for struggling readers in intervention schools were 8.39 points higher than spring scores for struggling readers in control schools ($b = 8.39, p < .01$). Again, there was evidence to that suggest that struggling readers who participated in TRI were beginning to catch up to their non-struggling peers. Struggling readers in the intervention schools made gains in Letter Word Identification skills that did not differ significantly from gains made by their non-struggling classroom peers ($b = -2.38, p = .15$). On the contrary, spring performance for struggling readers in control schools was lower than their non-struggling classmates' performance ($b = -8.36, p < .0001$). There was no evidence that non-struggling students in experimental schools underperformed relative to non-struggling students in control schools ($b = 2.41, p = .43$). Above and beyond intervention effects,

male students had lower Spring LW scores, first graders had higher spring LW scores, and higher maternal education was associated with higher spring scores.

Passage Comprehension. Controlling for differences in pre-test scores, TRI had a positive effect on struggling readers PC skills. Spring scores for struggling readers in intervention schools were approximately 7.65 points higher than spring scores for struggling readers in control schools ($p = .008$). However, there was not strong evidence that TRI promoted catch-up. Struggling readers in intervention schools made less gain than their non-struggling classmates did ($b = -8.29, p < .0001$), and struggling readers in control schools made less gain than their non-struggling classmates did ($b = -11.30, p < .0001$). Above and beyond intervention effects, students with higher maternal education, white students, and first graders had higher spring PC scores.

Spelling of Sounds. Spring scores for struggling readers in intervention schools were 3.23 points higher than spring scores for

struggling readers in control schools ($p = .02$). Struggling readers in intervention schools gained at the same rate as their non-struggling classmates as evidenced by a non-significant difference in spring performance ($b = 0.12$, $p = .88$). On the contrary, spring scores for struggling readers in control schools were lower than those of their non-struggling classmates ($b = -2.25$, $p = .03$). There was no evidence that non-struggling readers in intervention schools underperformed relative to non-struggling readers in control schools ($b = 0.87$, $p = .49$). Above and beyond intervention effects, higher maternal education was associated with higher spring Spelling of Sounds (SS) scores, and first graders had higher spring scores.

PPVT-III. There was no evidence that TRI had a positive effect on PPVT-III skills ($b = -1.59$, $p = .46$). There was evidence that spring performance for struggling readers in control schools differed from non-struggling readers in those schools ($b = -2.38$, $p = .02$). There was no evidence that non-struggling readers in intervention schools underperformed relative to non-struggling readers in control schools ($b = -1.80$, $p = .39$). Above and beyond intervention effects, white students had higher spring PPVT-III scores.

Discussion

The results from this study, using webcam technology to coach classroom teachers to individualize reading instruction for struggling readers, suggested that the TRI can significantly help struggling readers progress more quickly across a broad range of reading skills, including basic word reading, spelling, and passage comprehension skills over 1 year in comparison to children who did not receive this intervention. Furthermore, there was evidence that across some of these achievement measures in word reading and spelling of sounds the children in the TRI experimental group were able to progress at the same rate as their non-struggling peers. On the other hand, there was no evidence that the TRI could eliminate the gap between struggling and non-struggling readers in reading or improve receptive vocabulary over a one year period.

This study was important in a number of ways. First, webcam technology appeared to be an effective and efficient method to deliver professional development to remote rural schools using webcam coaching for classroom teachers. This study is one of only a few that has used technology to deliver professional development to classroom teachers (Mashburn et al., 2010; Powell et al., 2010) and the only one to use live webcam literacy coaching for classroom teachers that provided live and immediate feedback on teaching practices. In previous recent studies, teachers were given feedback on their teaching practices through delayed feedback from literacy consultants viewing videotapes of teachers' instructional practices. Moreover, as we have mentioned before, the live webcam sessions where the coach could see and hear the teacher working with a struggling reader and give real-time feedback were also likely critical in helping the teacher implement the intervention more quickly. In addition, the use of webcam technology allowed the teacher to have control of when the sessions took place and created efficiencies for the time allotted for the coaching sessions. For the current study, a half-time doctoral student could coach up to 12 classroom teachers at a time, given the flexibility afforded by the use of this technology. Although there was no cost/benefit analysis attempted in this study, there is no doubt that

the webcam technology was very affordable. Laptop computers were inexpensive at about \$800 a piece, and iChat and Skype were free to the users. The webcam technology could be implemented in almost any school since even remote rural schools have adequate Internet access. Thus, webcam technology as a tool for professional development builds on the previous work of others who have used videotape technology to provide professional development (Mashburn et al., 2010; Pianta et al., 2008; Powell et al., 2010).

Second, the TRI produced larger effect sizes than other studies, especially given that the classroom teacher was the one implementing the intervention. Across a broad range of reading assessments, the TRI produced effect sizes from .36 to .63 for kindergarten and first grade children over a 1-year period compared to children who did not receive the intervention. The current study had strong effects on both word level reading and reading comprehension, with a strong effect size of .48 on reading comprehension, even after accounting for school and classroom variance as well as maternal education, gender, and race. In addition, on two of the four reading measures, the TRI experimental children progressed at the same rate of growth in reading as the non-struggling readers in the same classrooms. Most other successful reading programs have been able to improve word reading skills over 1 year but many found no or small effects on reading comprehension (Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Torgesen et al., 1999). For instance, in a review of 42 one-on-one early reading intervention programs for at risk students, it was found that the effect sizes for word reading and passage reading were the greatest, with effect sizes from .41 to .54, whereas the effect sizes for reading comprehension were a modest .28 (Elbaum, Vaughn, Hughes, & Moody, 2000). In the current study, the effect sizes for word reading skills were .36 and .54, and the effect size for spelling was .63, certainly comparable to other studies. However, the effect size for reading comprehension in this study was almost double the average reported by Elbaum et al. (2000) across 42 intervention studies. This comprehension effect was probably due to the greater emphasis placed on reading comprehension compared with many other early reading interventions that often focus on improving children's decoding skills. The TRI emphasized not only word reading skills but comprehension skills during both *Word Work* and *Guided Oral Reading*, which made sure children could define the words they were reading, summarize stories they read, and answer complex questions about the texts, including causal and prediction questions.

In addition, previous studies that have used technology to help the classroom teacher, using video feedback on practices, have found positive effects, but the effect sizes in these studies were considerably smaller than reported in the current study. For instance, the most recent studies, using video feedback on language and literacy practices to help teachers in preschool (Mashburn et al., 2010; Powell et al., 2010), reported effect sizes for the children in the study of .1–.29. Although these were significant and important, the effect sizes in the current study for both word level reading and reading comprehension were .31–.63. These findings may suggest that future work consider webcam technology for feedback to teachers as the most effective way to get gains in reading for struggling readers.

Third, and a particularly important finding from this study, was the fact that the classroom teachers implemented successfully an

intervention for struggling readers with the help of literacy coaches. Most successful reading interventions for struggling readers have either used a specialized teacher to deliver the intervention outside the regular classroom or they have found effects only on word level reading (Foorman et al., 1998; Hurry & Sylva, 2007; Torgesen et al., 2001, 1999). This method of coaching classroom teachers appeared to be just as effective in helping struggling readers as employing one-on-one tutors. Elbaum et al. (2000) reported effect sizes for 42 tutoring interventions, with an average effect size of .41, which is comparable to the results in this study. Previous interventions using classroom teachers have not been particularly effective in helping struggling readers in early elementary school as suggested by reviews of the literature (Risko et al., 2008). Thus, this study is somewhat unique in not only demonstrating that the classroom teacher can implement effective instruction for struggling readers but that the students gain on a broad range of reading measures, including both word level reading skills and reading comprehension. Although we do not have direct evidence from this study, we believe, like other studies have argued, that literacy coaching (Carlisle & Berebitsky, 2011; International Reading Association, 2004) allowed teachers to get real time feedback on individualizing instruction for particular struggling readers (Scanlon, Gelzheiser, Vellutino, Schatschneider, & Sweeney, 2008; Speece, Case, & Molloy, 2003) that in turn enabled the children to gain in early reading.

Fourth, the teachers in this study were able to implement the TRI literacy strategies with relatively little training and with relatively modest instructional time per student. On average, teachers worked individually with a child two to three times per week for 6 weeks, with an average of 14 sessions for each child over the course of the year. In programs like "Reading Recovery," which used a specialized teacher, both more sessions and longer sessions were needed to achieve rapid progress (Elbaum et al., 2000; Schwartz, 2005). Even though our study used fewer resources and less time with individual children, the effect sizes for this study were comparable to those reported for a host of studies reviewed by Elbaum et al. (2000), all of which used a specialized teacher to deliver one-on-one intervention. Given the fewer resources available in low-wealth rural schools, the possibility of using the classroom teacher as the vehicle to help prevent reading failure in struggling readers and doing so with not much time taken away from the instructional day may have many benefits that hopefully can be replicated in future studies.

Limitations

There are a number of limitations in this study. First, there was a small school that was dropped from the study because of problems with technology. Even though there were only two classrooms in this school, this lack of participation of the school compromises our ability to make strong causal inferences. Second, there was limited information on fidelity. We were able to document exposure and adherence of the implementation (O'Donnell, 2008), but we were not able to measure the actual quality of the implementation. In future studies, it will be important to objectively observe the biweekly coaching sessions to more carefully document the quality of implementation beyond amount and adherence. This can now be done with the new capabilities to digitally record the iChat or Skype sessions. We believe that the

webcam coaching was the most important aspect of the intervention that led to change in student reading because of previous research that has demonstrated the importance of coaching (Al Otaiba et al., 2011; Carlisle & Berebitsky, 2011; Connor et al., 2011, 2009), but since we did not have a study that separated the effect of the summer institute from the coaching, we can only speculate on the reasons for the TRI success. We do know from previous work that workshops and institutes do not seem to be enough to produce real change in teachers that leads to improved reading for children (Garet et al., 2008; McGill-Franzen et al., 1999). Last, the TRI was not implemented long enough or intensely enough to allow the struggling readers to catch up with their non-struggling peers. Although the struggling readers were able to gain at the same rate as their non-struggling peers in three areas of early literacy when they received the TRI, the program was not able to allow the struggling readers to catch up with their non-struggling peers. It appears that future efforts may need programs, like the TRI, to be implemented more often for each child over 1 year, or for struggling readers to be involved in the TRI over multiple years, to make sure most struggling readers can catch up to their non-struggling peers.

Summary

Even with these limitations, this study is one of the first to suggest that the regular classroom teacher can learn effective instructional strategies using webcam literacy coaching that can lead to significant early reading gains in struggling readers and hopefully prevent reading failure in subsequent grades. It appears that efficient webcam technology may have contributed to the effectiveness of TRI by providing the regular classroom teacher with easy access to live feedback in the regular classroom on an ongoing basis over the school year. This webcam technology may be particularly effective in delivering professional development to classroom teachers in rural schools because these teachers do not have easy access to professional development opportunities due to geographic isolation. Webcam coaching may also be effective not only for rural schools but a wide variety of schools and could also be used to deliver professional development in other content areas for improving the instruction of classroom teachers.

References

- Al Otaiba, S., Connor, C. M., Folsom, J. S., Greulich, L., Meadows, J., & Li, Z. (2011). Assessment data-informed guidance to individualize kindergarten reading instruction: Findings from a cluster-randomized control field trial. *The Elementary School Journal*, 111, 535-560. doi: 10.1086/659031
- Al Otaiba, S., Hosp, J. L., Smartt, S., & Dole, J. A. (2008). The challenging role of a reading coach, a cautionary tale. *Journal of Educational & Psychological Consultation*, 18, 124-155. doi:10.1080/10474410802022423
- Amendum, S., Vernon-Feagans, L., & Ginsberg, M. (2011). The effectiveness of a technologically facilitated classroom-based early reading intervention. *Elementary School Journal*, 112, 107-131. doi:10.1086/660684
- Bear, D. R., Invernizzi, M., Templeton, S., & Johnston, F. (2003). *Words their way: Word study for phonics, vocabulary, and spelling instruction* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Beck, I. L. (2006). *Making sense of phonics: The hows and whys*. New York, NY: Guilford Press.

- Blok, H., Oostdam, R., Otter, M. E., & Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: A review. *Review of Educational Research*, 72, 101–130. doi:10.3102/00346543072001101
- Brooks-Gunn, J., & Duncan, G. J. (1997). The effects of poverty on children. *The Future of Children*, 7, 55–71. doi:10.2307/1602387
- Carlisle, J. F., & Berebitsky, D. (2011). Literacy coaching as a component of professional development. *Reading and Writing*, 24, 773–800. doi:10.1007/s11145-009-9224-4
- Chambers, B., Slavin, R. E., Madden, N. A., Abrami, P., Logan, M. K., & Gifford, R. (2011). Small-group, computer-assisted tutoring to improve reading outcomes for struggling first and second graders. *The Elementary School Journal*, 111, 625–640. doi:10.1086/659035
- Cole, J., & Hilliard, V. (2006). The effects of web-based reading curriculum on children's reading performance and motivation. *Journal of Educational Computing Research*, 34, 353–380. doi:10.2190/H43W-1N3U-027J-07V5
- Connor, C., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., . . . Schatschneider, C. (2011). Testing the impact of child characteristics \times instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly*, 46, 189–221.
- Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007, January 26). The early years: Algorithm-guided individualized reading instruction. *Science*, 315, 464–465. doi:10.1126/science.1134513
- Connor, C., Morrison, F. J., & Petrella, J. N. (2004). Effective reading comprehension instruction: Examining child \times instruction interactions. *Journal of Educational Psychology*, 96, 682–698. doi:10.1037/0022-0663.96.4.682
- Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., . . . Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child by instruction interactions on first graders' literacy development. *Child Development*, 80, 77–100. doi:10.1111/j.1467-8624.2008.01247.x
- Deweese, S. (2000). *Participation of rural schools in Comprehensive School Reform Demonstration Program: What do we know?* Washington, DC: Office of Educational Research and Improvement.
- Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test—III*. Circle Pines, MN: American Guidance Service.
- Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology*, 92, 605–619. doi:10.1037/0022-0663.92.4.605
- Elish-Piper, L., & L'Allier, S. K. (2011). Examining the relationship between literacy coaching and student reading gains in grades K–3. *The Elementary School Journal*, 112, 83–106. doi:10.1086/660685
- Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, 90, 37–55. doi:10.1037/0022-0663.90.1.37
- Franklin, C. (2007). Factors that influence elementary teachers use of computers. *Journal of Technology and Teacher Education (JTATE)*, 15, 267–293.
- Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., & Jones, W. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). Washington, DC: Institute of Education Sciences.
- Government Accountability Office. (2004). *No Child Left Behind Act: Additional assistance and research on effective strategies would help small rural districts* (Report GAO-04-909). Washington, DC: Author.
- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8, 206–213. doi:10.1007/s11121-007-0070-9
- Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: RAND.
- Haager, D., Klingner, J., & Vaughn, S. (Eds.). (2007). *Evidence-based reading practices for response to intervention*. Baltimore, MD: Brookes.
- Huffstetter, M., King, J. R., Onwuegbuzie, A. J., Schneider, J. J., & Powell-Smith, K. A. (2010). Effects of a computer-based early reading program on the early reading and oral language skills of at-risk pre-school children. *Journal of Education for Students Placed at Risk*, 15, 279–298. doi:10.1080/10824669.2010.532415
- Hurry, J., & Sylva, K. (2007). Long-term outcomes of early reading intervention. *Journal of Research in Reading*, 30, 227–248. doi:10.1111/j.1467-9817.2007.00338.x
- International Reading Association. (2004). *The role and qualifications of the reading coach in the United States*. Newark, DE: International Reading Association.
- Kennedy, M. J., & Deshler, D. D. (2010). Literacy instruction, technology, and students with learning disabilities: Research we have, research we need. *Learning Disability Quarterly*, 33, 289–298.
- Lee, V. E., & Burkham, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.
- Lewandowski, L., Begeny, J., & Rogers, C. (2006). Word-recognition training: Computer versus tutor. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, 22, 395–410. doi:10.1080/10573560500455786
- Lyon, G. R. (2001, March 8). *Measuring success: Using assessments and accountability to raise student achievement*. Statement presented at the Hearing before the Subcommittee on Education Reform, Committee on Education and the Workforce, U.S. House of Representatives, Washington, DC.
- Mashburn, A. J., Downer, J. T., Hamre, B. K., Justice, L. M., & Pianta, R. C. (2010). Consultation for teachers and children's language and literacy development during pre-kindergarten. *Applied Developmental Science*, 14, 179–196. doi:10.1080/10888691.2010.516187
- Mathes, P. G., Torgesen, J. K., & Allor, J. H. (2001). The effects of Peer Assisted Learning Strategies for First Grade Readers with and without additional computer assisted instruction in phonological awareness. *American Educational Research Journal*, 38, 371–410. doi:10.3102/00028312038002371
- McGill-Franzen, A., Allington, R. L., Yokoi, L., & Brooks, G. (1999). Putting books in the classroom seems necessary but not sufficient. *The Journal of Educational Research*, 93, 67–74. doi:10.1080/00220679909597631
- McKenna, M. C., & Walpole, S. (2008). *The literacy coaching challenge: Models and methods for grades K-8*. New York, NY: Guilford Press.
- Morgan, D. N., Timmons, B., & Shaheen, M. (2006). Tutoring: A personal and professional space for preservice teachers to learn about literacy instruction. In J. V. Hoffman, D. L. Shcallert, C. M. Fairbanks, J. Worthy, & B. Maloch (Eds.), *55th yearbook of the National Reading Conference* (pp. 212–223). Oak Creek, WI: National Reading Conference.
- Morris, D., Tyner, B., & Perney, J. (2000). Early Steps: Replicating the effects of a first-grade reading intervention program. *Journal of Educational Psychology*, 92, 681–693. doi:10.1037/0022-0663.92.4.681
- National Center for Education Statistics. (2009). *The nation's report card: Reading 2009* (NCES 2010–458). Washington, DC: Institute of Education Sciences, U.S. Department of Education.
- O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, 78, 33–84. doi:10.3102/0034654307313793

- O'Hare, W. P. (2009). *The forgotten fifth: Child poverty in rural America*. Durham, NH: The Carsey Institute.
- Pianta, R. C., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, 23, 431-451. doi:10.1016/j.ecresq.2008.02.001
- Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on head start teachers and children. *Journal of Educational Psychology*, 102, 299-312. doi:10.1037/a0017763
- Provasnik, S., KewalRamani, A., Coleman, M. M., Gilbertson, L., Herring, W., & Xie, Q. (2007). *Status of education in rural America* (NCES 2007-040). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.
- Risko, V. J., Roller, C. M., Cummins, C., Bean, R. M., Block, C. C., Anders, P. L., & Flood, J. (2008). A critical analysis of research on reading teacher education. *Reading Research Quarterly*, 43, 252-288. doi:10.1598/RRQ.43.3.3
- Saine, N. L., Lerkkanen, M., Ahonen, T., Tolvanen, A., & Lyytinen, H. (2011). Computer-assisted remedial reading intervention for school beginners at risk for reading disability. *Child Development*, 82, 1013-1028. doi:10.1111/j.1467-8624.2011.01580.x
- Scanlon, D. M., Gelzheiser, L. M., Vellutino, F. R., Schatschneider, C., & Sweeney, J. M. (2008). Reducing the incidence of early reading difficulties: Professional development for classroom teachers versus direct interventions for children. *Learning and Individual Differences*, 18, 346-359. doi:10.1016/j.lindif.2008.05.002
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147-177. doi:10.1037/1082-989X.7.2.147
- Schwartz, R. M. (2005). Literacy learning of at-risk first-grade students in the Reading Recovery early intervention. *Journal of Educational Psychology*, 97, 257-267. doi:10.1037/0022-0663.97.2.257
- Speece, D. L., Case, L. P., & Molloy, D. W. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research & Practice*, 18, 147-156. doi:10.1111/1540-5826.00071
- Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K., & Conway, T. (2001). Intensive remedial instruction for students with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities*, 34, 33-58. doi:10.1177/002221940103400104
- Torgesen, J., Wagner, R., Rashotte, C., Lindamood, P., Rose, E., Conway, T., & Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology*, 91, 579-593. doi:10.1037/0022-0663.91.4.579
- Vernon-Feagans, L., Gallagher, K. C., & Kainz, K. (2010). The transition to school in rural America: A focus on literacy. In J. L. Meece & J. S. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 163-184). New York, NY: Erlbaum.
- Vernon-Feagans, L., Kainz, K., Amend, S., Ginsberg, M., Wood, T., & Bock, A. (2012). Targeted Reading Intervention: A coaching model to help classroom teachers with struggling readers. *Learning Disability Quarterly*, 35, 102-114. doi:10.1177/0731948711434048
- Woodcock, R. W., Mather, N., & Schrank, F. A. (2004). *Woodcock-Johnson III Diagnostic Reading Battery*. Itasca, IL: Riverside Publishing.

Received December 7, 2011

Revision received September 25, 2012

Accepted December 18, 2012 ■

Using Electronic Portfolios to Foster Literacy and Self-Regulated Learning Skills in Elementary Students

Philip C. Abrami, Vivek Venkatesh, Elizabeth J. Meyer, and C. Anne Wade
Centre for the Study of Learning and Performance, Concordia University

The research presented here is a continuation of a line of inquiry that explores the impacts of an electronic portfolio software called ePEARL, which is a knowledge tool designed to support the key phases of self-regulated learning (SRL)—forethought, performance, and self-reflection—and promote student learning. Participants in this study were 21 teachers from elementary schools (Grades 4–6) and their students ($N = 319$) from 9 urban and rural English school boards in Quebec and Alberta, Canada, who participated during the 2008–2009 school year. Students with low enthusiasm for the use of ePEARL were excluded from the main sample as they exhibited different patterns in learning gains and self-regulatory skills as compared with those with high and medium enthusiasm. Multivariate analyses of covariance showed that students motivated to use the software made significantly greater gains compared with controls in 3 of 4 writing and reading skills ($p < .01$) as assessed by the constructed response subtest of the Canadian Achievement Test (fourth edition). Multivariate analyses of covariance of student survey data revealed that, over time, students who used the software reported higher levels of SRL processes than those in the control group ($p < .01$). Implications of the findings for school leaders and teacher educators regarding the use of electronic portfolios are discussed.

Keywords: self-regulation, electronic portfolios, quasi-experimental research, elementary education, classroom research

A recent international report on the performance of secondary-school students in Western, industrialized countries (Knighton, Brochu, Gluszynski, 2010; Organization for Economic Co-operation and Development, 2010) found that a significant number of students in every country lacked fundamental literacy, numeracy, and scientific reasoning skills. In addition, some students may lack the sophisticated strategies for learning how to learn, strategies that may be increasingly important in the knowledge age. Furthermore, these gaps in essential competencies and skills have substantial personal, social, and economic consequences. For example, Statistics Canada (Coulombe, Tremblay, & Marchand, 2004) estimated that a 1% increase in the rate of adult literacy in the Canadian population of about 30 million inhabitants would be worth \$18.4 billion dollars annually to the Canadian economy.

Contemporary trends in education research indicate that improvements in educational success will occur when students be-

come more active, engaged participants in their learning, enhancing the extent to which learning is personally meaningful (e.g., Tobias & Duffy, 2009). These trends also recognize that lifelong learning skills, or the ability to develop and use learning strategies and skills, are growing increasingly important in the knowledge age where content expertise is evolving rapidly. Put colloquially, the importance of knowing how is supplanting the importance of knowing what.

Self-Regulated Learning

Against this backdrop, there is increasing interest in theories and research on student-centered learning and their application to classroom practice. Self-regulated learners are individuals who are metacognitively, motivationally, and behaviorally active participants in their own learning (Zimmerman, 2000; Zimmerman & Schunk, 2011). Zimmerman (2000) defines self-regulation as “self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals” (p. 14). This implies not only behavioral skill management and subject knowledge but also metacognitive awareness, social influences, and motivational beliefs about personal agency. Zimmerman structures the self-regulation process in three phases: *forethought*, *performance*, and *self-reflection*. The *forethought* phase includes task analysis in the form of goal setting and strategic planning, and self-motivation beliefs in the form of self-efficacy, outcome expectations, intrinsic interest, and goal orientation. The *performance* phase is divided into *self-control*, which includes self-instruction, imagery, attention-focusing and task strategies, and *self-observation*, which includes self-recording and self-experimentation. The third phase is *self-reflection*. It includes

This article was published Online First September 9, 2013.

Philip C. Abrami, Vivek Venkatesh, Elizabeth J. Meyer, and C. Anne Wade, Centre for the Study of Learning and Performance, Concordia University, Montreal, Quebec, Canada.

Elizabeth J. Meyer is now at the School of Education, California Polytechnic State University.

This research was supported by grants from the Social Sciences and Humanities Research Council of Canada, the Canadian Council on Learning, and the *Fonds de recherche sur la société et la culture*.

Correspondence concerning this article should be addressed to Philip C. Abrami, Centre for the Study of Learning and Performance, Concordia University, 1455 DeMaisonneuve Boulevard West, Montreal, Quebec, Canada H3G 1M8. E-mail: abrami@education.concordia.ca

self-judgment, composed of self-evaluation and causal attribution, as well as self-reaction, which involves self-satisfaction and adaptive/defensive responses. The process is described as cyclical because successful self-regulation depends on the constant monitoring and correction of performance based on feedback about recent efforts.

Zimmerman's (2000) self-regulation model is a social-cognitive model that puts great emphasis on social, environmental, and personal influences in efforts to self-regulate effectively. In this model, using valuable social resources (such as peer or teacher feedback, modeling and emulation of expert behavior) or environmental support (such as self-rewarding achievement with a relaxing activity) will result in successful self-regulation.

There are numerous studies of the effectiveness of self-regulated learning (SRL) on academic achievement (e.g., Chung, 2000; Paris & Paris, 2001; Winne, 1995; Zimmerman, 1990; Zimmerman & Bandura, 1994; Zimmerman & Martinez-Pons, 1988) as well as on learning motivation (Pintrich, 1999). Furthermore, SRL is considered by some as a key competence for lifelong learning (European Union Council, 2002). Considering these three areas—academic performance, motivation to learn, and learning strategies—where students can benefit from SRL, the potential value of SRL training programs becomes clear. Providing students with knowledge and skills about how to self-regulate their learning may help them to self-initiate motivational, behavioral, and metacognitive activities in order to control their learning (Zimmerman, 1998). Recent empirical work by Venkatesh and Shaikh (2008, 2011) and Shaikh, Zuberi, and Venkatesh (2012) has also demonstrated links between academic performance and specific self-regulatory processes, namely, task understanding and monitoring proficiencies. Our work takes into account these relationships by including both academic achievement and self-regulation measures in the analyses.

In two related meta-analyses, Dignath and Buettner (2008) and Dignath, Buettner, and Langfeldt (2008) investigated the impact of various SRL training characteristics on academic performance, strategy use, and the motivation of students. The meta-analyses included 49 studies conducted with primary-school students and 35 studies conducted with secondary-school students. Altogether, 357 effect sizes were analyzed.

For achievement outcomes, the average effect size for SRL training was + 0.61 for primary schools and + 0.51 for secondary schools. For cognitive and meta-cognitive strategy use, the average effect size for primary schools was + 0.72 and + 0.88 for secondary schools. For motivation outcomes, the average effect size for primary schools was + 0.75 and + 0.17 for secondary schools. For both school levels, effect sizes were higher when researchers conducted the training instead of regular teachers. Moreover, interventions attained higher effects when mathematics was the subject matter rather than in reading/writing or other subjects. The reviewers concluded that SRL can be fostered effectively at both primary- and secondary-school levels. The current research explores whether a technology-based tool for fostering SRL would also have positive impacts.

Knowledge Tools and Electronic Portfolios

The potential for technology to radically transform and improve education is widely recognized by policy makers, scholars, and practitioners (Campuzano, Dynarski, Agodini, Rall, & Pendleton,

2009; Canadian Council on Learning, 2008; CEO Forum on Education and Technology, 2001; Dynarski et al., 2007; Ungerleider & Burns, 2003; Zimmerman & Tsikalas, 2005); however, there have been mixed results when new technologies meet the realities of the diverse and changing classroom contexts of schools (Abrami et al., 2006; Abrami, Savage, Wade, & Hipps, 2008; Avramidou & Zembal-Saul, 2003; Azevedo, 2005; Barrett, 2007; Bernard, Bethel, Abrami, & Wade, 2007; Cuban, 1993; Cuban, Kirkpatrick, & Peck, 2001). Explanations vary including the lack of technology infrastructure, insufficient training and support of teachers (Meyer, Abrami, Wade, Aslan, & Deault, 2010), and knowledge tools that may lack key design principles guided by what has been learned from the learning and motivational sciences (Abrami, 2010; Abrami, Bernard, Bures, Borokhovski, & Tamim, 2011; Pintrich, 2003; Mayer, 2001, 2008; Wozney, Venkatesh, & Abrami, 2006). The current development and research project attempts to overcome some of these challenges by designing software that is both faithful to Zimmerman's (2000) social-cognitive model of self-regulation and by providing both help and support embedded in the software as well as face-to-face training and support leading to enhanced implementation fidelity.

The research presented here is a continuation of a line of inquiry that explores the impacts of one particular knowledge tool—an electronic portfolio (EP). EPs build on the evidence of what is already known about effective portfolio pedagogy, and makes working with portfolios more engaging, dynamic, and accessible for students, teachers, and parents. An EP is a digital container capable of storing visual and auditory content, including text, images, video, and sound. EPs may also be learning tools not only because they organize content but also because they are designed to support a variety of pedagogical processes and assessment purposes. Historically speaking, EPs are the knowledge age's version of the artist's portfolio for students in the sense that they not only summarize a student's creative achievements but also illustrate the process of reaching those achievements. An artist, architect, engineer, or student who displays her or his portfolio of work allows the viewer to form a direct impression of that work without having to rely on the judgments of others. EPs tell a story both literally and figuratively by keeping a temporal and structural record of events. EPs can offer valuable opportunities for integrating technology into K–12 classrooms beyond serving as multimedia containers. They may serve to deepen a student's learning experiences by placing the student at the center of his or her learning and scaffolding essential meta-cognitive skills such as goal setting, identifying strategies, and reflecting on one's learning (Abrami & Barrett, 2005).

According to Abrami and Barrett (2005), EPs have three broad purposes: process, showcase, and assessment. All three types of EPs can be used to display selected artifacts and would enable learners to develop their metacognitive skills in choosing their pieces, reflecting on how they meet assessment criteria and re-working elements on the basis of feedback. We chose to work with EPs designed as *process portfolios* that support how users learn through embedded structures and strategies. A *process EP* can be defined as a purposeful collection of student work that tells the story of a student's effort, progress, and/or achievement in one or more areas (Arter & Spandel, 1992; Barrett, 2007; MacIsaac & Jackson, 1994). Process portfolios are personal learning management tools. They are meant to encourage individual improvement,

personal growth and development, and a commitment to lifelong learning. The authors are especially interested in the use of EPs as process portfolios to support learning.

Process EPs are gaining in popularity for multiple reasons (Abrami & Barrett, 2005; Barrett, 2009; Zubizarreta, 2004). They provide multimedia display and assessment possibilities for school and work contexts, allowing the use of a variety of means to demonstrate and develop understanding—especially advantageous for children whose competencies may be better reflected through these authentic tasks (Barrett, 2007). At the same time, by engaging learners, their deficiencies in core competencies may be better overcome. Process EPs may scaffold attempts at knowledge construction by supporting reflection, refinement, conferencing, and other processes of self-regulation, important skills for lifelong learning and learning how to learn. They are useful for cataloging and organizing learning materials, readily illustrating the process of learner development. They can also provide remote access encouraging anywhere, anytime learning and easier input from peers, parents, and teachers (Barrett, 2008).

Process EPs are linked to students' abilities to self-regulate their learning and to enhance their development of important educational skills and abilities, especially literacy skills (Meyer et al., 2010; Wade, Abrami, & Sclater, 2005). When students use portfolios, they assume more responsibility for their learning, better understand their strengths and limitations, and learn to set goals (Hillyer & Lye, 1996).

Unfortunately, evidence to date on the impacts of EPs on learning and achievement is sparse (Barrett, 2007; Carney, 2005; Zeichner & Wray, 2001). Some research has filled this gap by studying the impact of EPs on teaching and learning processes, especially those related to self-regulation, in late elementary classrooms. Abrami et al. (2008) found that K–12 teachers faced some challenges when attempting to integrate process EPs into their teaching; hence, the levels of use throughout the school year were fairly low. Although the teachers had a positive view of EPs and SRL processes, most teachers needed to adjust their teaching strategies to effectively incorporate EPs into their teaching and needed time and support to do so.

Meyer et al. (2010) studied high and medium EP implementing experimental teachers, who had been trained and supported, plus nonimplementing control teachers and their students. In this quasi-experiment, the researchers found that older elementary students who were in classrooms where the teacher provided regular and appropriate use of the EP tool, compared with control students who did not use the tool, showed significant improvements in their writing skills on a standardized literacy measure and certain metacognitive skills, such as monitoring, measured via student self-report. This is the first study to offer some evidence that teaching with EPs had positive impacts on students' literacy achievement and SRL skills when the tool was used regularly and integrated into classroom instruction.

Meyer, Abrami, Wade, and Scherzer (2011) collected data to understand how teachers used EPs in their classrooms, to what extent they integrated the EP into their practice, and the factors that influenced their use. They found that low implementers experienced significant technical obstacles and/or were reluctant to change their established practices, whereas high implementers reported feeling supported by their administration and experienced

growth in their teaching practice as a result of the scaffolding and support provided by the software.

On the basis of these findings, Meyer et al. (2011) made several recommendations to increase the quality of EP use. First, it is essential to ensure that the classrooms and schools have sufficient technical infrastructure to support the innovation. Second, teachers and students must have consistent access to functioning computers that are regularly maintained. Third, teachers must feel that there is positive support from the administration to invest the time in learning to teach with EPs. Otherwise, teacher training needs to focus more on why using educational technology such as EPs is important and appropriate rather than only how to use EPs. Most EPs are not technically difficult tools to use, but they may be pedagogically challenging. Some EPs focus on student-centered learning, which means that teachers need to accept classroom practices that go beyond didactic forms of instruction. The last suggestion is to provide external encouragement for teachers to adapt an innovation and provide a culture that values experimentation, improvement, and evidence-based practices.

One purpose of the current research was to determine whether the findings of Meyer et al. (2010, 2011) could be meaningfully extended. By ensuring better access to technology, enhanced support and recognition from school administrators, and better and more professional development both embedded in the tool (e.g., just-in-time instructional video vignettes) and via training and follow-up support, it was hoped that the number of high-EP implementers would increase and that important learning outcomes would be evident. Dignath and Buettner (2008) and Dignath et al. (2008) found that researcher implementations of SRL training have larger effects than teacher implementations. But in the current investigation, teacher implementations were chosen as a means to foster authentic classroom practices where teachers and their students used the software for extended periods to teach and learn required curricular content. It was hoped that these would be scalable and sustainable practices, not researcher-made demonstrations, and that the ways to promote effective and long-lasting change in real classrooms would be better understood.

Abrami (2010; Abrami et al., 2011) commented critically on the widely held belief that learners are motivated to use knowledge tools for learning. First, learners may not value the outcome(s) of learning sufficiently to increase their efforts to learn—it is not so important to do well. Second, learners may believe that gains in learning from increased effort are inefficient—it takes too much effort to do a little bit better. Third, learners may not want to become more responsible for their own learning—it is too risky unless the perceived chances of a positive outcome are increased. Fourth, learners may believe that novel approaches to learning increase the likelihood of poor outcomes, not increase them—it is not of interest or too risky because they do not believe the tool will help them learn. Therefore, a secondary purpose of this research was to explore the extent of student engagement and satisfaction with the use of EPs.

Zimmerman (2008) published an overview of SRL research that studied several online tools designed to stimulate and study various SRL processes in students. He identified four key trends that remain as important questions, including the relationship between student reports of SRL and actual use of SRL processes, the relationship between levels of SRL and overall academic achievement, the role of the social context of the classroom in stimulating

or hindering the development of SRL skills, and finally the relationship between motivation and SRL processes. The research presented in this article addresses two of these four areas: the relationship between levels of SRL and overall academic achievement and the role of the social context of the classroom in stimulating or hindering the development of SRL skills.

Research Context and Method

About ePEARL

Zimmerman and Tsikalas' (2005) review of computer-based learning environments (CBLEs) designed to support SRL provides a framework for the development of a tool to support the three cyclical phases of SRL: forethought, performance, and self-reflection. The lessons learned from other partially SRL-supportive CBLEs has enabled us to plan for effective SRL-supportive design of an *Electronic Portfolio Encouraging Active and Reflective Learning* (ePEARL).

The Centre for the Study of Learning and Performance (Concordia University, Montreal, Canada), in collaboration with our partner LEARN (Montreal, Canada), developed ePEARL as a bilingual (English and French), web-based, student-centered EP software tool, which is designed to support the phases of self-regulation (Zimmerman, 2000; Zimmerman & Tsikalas, 2005).

The latest version of ePEARL may be explored by visiting <http://grover.concordia.ca/epearl/promo/en/index.php>. The slightly older version (3.0) used in this research is archived on our university server. The software is available at no cost to educators.

EP tasks involved in the forethought phase are setting outcome goals, setting process goals, documenting goal values, planning strategies, and creating learning logs. Tasks involved in the performance phase are creating work, self-examination through recordings and drafts of work, and learning log entries. Tasks involved in the self-reflection phase are reflecting on work, process, and feedback received and becoming aware of new goal opportunities.

Developed in Hypertext Preprocessor (Version 5.X) using a MySQL database, four levels of ePEARL were designed for use by children, teachers, and adult learners in early elementary, late elementary, secondary, and postsecondary schools and institutions. The forethought or planning phase includes the following features: describing the task, setting outcome and process goals, identifying strategies to achieve those goals, and providing a place for teachers to provide scaffolding and feedback. Figure 1 displays the planning phase for a science project for which the students worked in teams to build an "insulating machine" that would keep water hot. Students enter information under "Task Description," "Criteria," "Task Goals," and "Strategies"; these processes are often carefully

Insulating machine

Folder Science/technolo
Colour Code finish
Date 01/22/09
Teacher Colour Codes ? achieved competency

[printable version](#)

[My General Goals](#)

Task Description

Task Description

Make an Insulating machine

Criteria

The machine should keep the water hot

Goals

Task Goals Updated 04/22/09

1. Choose item to make a machine
2. Built the machine
3. write hypothesis
4. test the temperature before and after the test
5. test it
6. write the conclusion

Strategies Updated 04/22/09

1. We did research to choose the best heat perserving materials
2. Carefully measure the temperature
3. Cooperate with other groups
4. Write a logical hypothesis and conclusion



Teacher

Updated 04/21/09

Specific and therefore excellent!

Figure 1. Insulating Machine project: Planning.

scaffolded and modeled by the teachers. The teacher completes the “teacher color codes” and the yellow feedback box at the bottom.

Features available in the performance or doing phase include creating new work; linking to existing work; attaching digital files to document the work completed; and space for teacher feedback on the content. Figure 2 displays the content and the teacher’s feedback for the insulating machine assignment. Students composed the text seen in Figure 2 and uploaded the image, after which the teacher provided feedback.

Features available in the reflecting phase include reflecting on work; sharing work; obtaining feedback from teachers, peers, and parents; editing work; saving work under multiple versions; and sending work to a presentation folder. ePEARL promotes the creation of general learning goals for a term or year, or for a specific work/artifact; reflection; and peer, parent, and teacher feedback on the entire portfolio or on a specific artifact. Students and teachers can monitor individual progress toward completing each SRL component by viewing the ePEARL index page. Figure 3 presents a sample index page from a student portfolio. This index page displays the title of each artifact, the date modified, the number of file attachments (A), if goals have been set (G), if reflections have been completed (R), and if comments have been provided (C) by teachers (T) or students (S). The remaining columns (BAL, CCC, SA, CK) link to specific learning outcomes described by local government education policy.

ePEARL guides students through the creation process, allowing enough flexibility for truly creative work and just enough scaffolding to keep students on the right track. It offers a text editor and an audio recorder for the creation of work. Readings, music pieces, or oral presentations may be recorded. The software also offers the ability to attach work completed using other software, so it can accommodate any kind of digital work a student creates, including videos, slideshows, podcasts, scanned images, or photographs of paper-based work. See Figure 2.

Before work is created, students are encouraged to set goals for their work, and may attach learning logs, evaluation rubrics, and study plans to keep track of their learning process as it takes place. After the creation of work, sharing with peers or teachers is supported so that students may solicit feedback on drafts of work. Figure 4 shows the sharing screen, which allows students to select individual classmates or entire classes that they would like to give permission to in order to view and comment on their work. In this image, you also see the graphic of the SRL process that is present through the planning stages to help students remember each step of the process. Scaffolding for these processes is embedded in the portfolio, as seen in the text box that describes the purpose of “sharing” an artifact in order to get feedback. Students may also reflect on their performance and strategies and use these reflections to adjust their goals for the next work.

Content

Text

Were attaching a file with a picture of our insulating machine.

Files



CIMG1200.JPG

[view](#) | [download](#)

Feedback



Teacher Feedback ?

Teacher

Updated 04/21/09

You describe the task and goals with precision. The picture of your insulating machine is very good too. Please tell us what you learned from this experiment.

Fantastic!

Figure 2. Insulating Machine project: Content and teacher feedback.

Readers Theater

Folder reading

Colour Code Working on

Date 11/21/08

Teacher Colour Codes none

goals
reflection
feedback



Task Description

Task Description

We are doing Readers theater to help us read and so then we can also get rid of stage fright.

Criteria

1. I use different voices
2. The audience is enthused
3. Performance is smooth and lively
4. All words are understood
5. Everyone can hear
6. Performance is polished

Goals

Task Goals Updated 11/21/08

My goal I am also going to try to get different voices because I use my regular voice for everything. I am also hoping that I can get through the whole play without laughing. I am also hoping that I can understand the reading.

Strategies Updated 11/21/08

When I do Readers Theater: I will try and go to Julia or Mrs. [redacted] and ask for help on words I don't understand.

Figure 5. Readers theater: Planning.

"Task Goals" and "Strategies." In Figure 6, the student composed a reflection at the end of the project; the teacher and the student's peers could view the final artifact and enter their feedback.

In addition, there are both prose and extensive multimedia support materials for teachers and students to develop a better understanding of the what, why, and how of the self-regulation processes supported by ePEARL. The research team created a wiki and a virtual tutorial to help support teachers' implementation of the SRL features within ePEARL. A series of six "jump start" lessons were provided to members of the wiki to help teachers introduce ePEARL and various SRL processes such as setting general goals, organizing your work, setting task goals, and reflecting and providing feedback. An overview of this jump start program is provided in Appendix A. The virtual tutorial contains a series of 2-min videos showcasing the software and providing pedagogical supports and examples for classroom teachers. This tutorial can be accessed by visiting <http://grover.concordia.ca/epearl/tutorial/index.php>. Additionally, supports were embedded within the software through help buttons that both students and teachers could access that provided definitions of SRL terminology, sample responses, and hyperlinks to the virtual tutorial. The professional development and just-in-time materials support the demonstration and modeling of student-centered skills and instruction, explanations of those skills, and elaboration of skills through

additional support materials. In fact, the ePEARL project team went to some length to ensure that both students and teachers were knowledgeable about the features of ePEARL and the processes of SRL. This effort was reflected not only in the embedded multimedia and prose support that are part of the tool but also because of the training and follow-up support the team provided throughout the project.

Study Design

Participants in this study were 21 teachers from elementary schools (Grades 4–6) and their students ($N = 319$) from nine urban and rural English school boards in Quebec and Alberta, Canada, who participated during the 2008–2009 school year. All experimental teachers ($n = 9$) received at least a half-day of training on the use of ePEARL from research center staff and follow-up support, including lesson plans and job aids, an online discussion forum (in the form of a moderated wiki), as well as in-class observations and model lessons during the school year. In addition, multimedia scaffolding and support for teachers and students are embedded in the tool.

School principals and school board administrators were consulted to identify control teachers and their classrooms that would match as closely as possible the experimental teachers and their

Motivation



Reflections

Reflections Updated 01/12/09

I think I did pretty well I've met all of my goals, I think I could have done better because I don't really look at the crowd because I get nervous, but the bad thing is that not many people could hear what I was saying, most of the people were going haa, what is she saying? When I look at the crowd I do a really weird smile, kind of like I'm weird too, Any idea's?

A couple of people said to try it out on crowds before we share with the classes. I think that my friends are really great with giving me advice. I will share everything with my friends because they can help me with everything. I'm glad I know them Mrs. said that I should meet my goals and try and have fun with your Readers Theater and see what happens, JUST HAVE FUN!!

Edit Comments updated

Save

Feedback



Teacher Feedback

Teacher

Updated 11/23/08

You have a lot of goals and strategies. They are all good so let's see if we can work together to reach them. They would make a difference to all your reading skills and your presentation skills.

Edit Feedback



Peer Feedback

Updated 01/14/09

I really liked your play because it was really funny. I think it could be a little louder. Try reading it to your brother. Can he hear you?

Figure 6. Readers theater: Reflecting and feedback.

classrooms. All teachers needed to follow provincial curriculum requirements for the development of language arts skills. Experimental teachers did so with the aid of the software, whereas control teachers did not. All teachers were at liberty to decide on the provision of the type of language arts instruction. There were no special language arts materials provided to either experimental or control teachers by the research team. Given there was little incentive to participate, control teachers were offered a stipend that could be used toward the purchase of classroom resources or professional development. Informed consent was obtained from students' parents following Canada's Tri-Council Policy on the ethical treatment of research participants.

The study used multivariate and univariate analyses of covariance to identify differences between the experimental and control groups on measures of reading, writing, and self-regulation. Teacher and student questionnaire data on self-regulation were collected in September and October of 2008. Teacher and student questionnaire data were collected again in May and June of 2009 after the software was used for some part of the school year,

ranging from 6 to 8 months. In addition to questionnaires, all students completed the constructed response subtest of the Canadian Achievement Test, fourth edition (CAT-4; Canadian Test Centre, 2008) in both the fall and spring to assess their reading and writing skills.

Instrumentation

The CAT-4 assesses both response to text (ideas, support) and writing (content, content management) using a rubric. The results of the CAT-4-constructed response reading and writing activities were sent to the Canadian Test Centre for evaluation and as part of their norming study. They assigned final scores to all the students that were then mailed back for inclusion in the data set. The constructed response subtest depends on student narrative responses to prompts as opposed to the multiple-choice format of the main tests of the CAT-4, which also measures student literacy. Multiple story prompts were used in each class at both pretest and posttest, but no student responded to the same prompt twice. This

form of measuring literacy achievement was used because it was compatible with notions of authentic assessment, even though it meant we generated a less detailed analysis of student learning than using the closed-ended version of the CAT-4. The reliability coefficients (Kuder-Richardson 20) for CAT-4 subtests range between 0.85 and 0.95, depending on the level and subtest. In the previous version (CAT-3), test validity was established by showing that grade levels that were known to have different levels of achievement did indeed have different mean scores on the same test.

Abrami and Aslan (2007) developed the Student Learning Strategies Questionnaire (SLSQ; see also Abrami et al., 2008, and Appendix B) used in the current and past studies to measure students' perceptions of their use of SRL strategies, including their ability to set learning goals, observe and correct their performance, and reflect on the learning outcome. The SLSQ contains six scales, namely, Goal Setting, Strategy Planning, Self-Observation, Self-Instruction, Feedback from Adults, and Self-Evaluation. In addition, students were asked to complete the Academic Self-Regulation Questionnaire (SRQ; Ryan & Connell, 1989). This 32-item instrument measures four subscales of self-regulation, namely, External, Identified, Intrinsic, and Introjected. The composite score of the SRQ, known as the Relative Autonomy Index, is calculated from the four subscales as follows: $2 \times \text{intrinsic} + \text{identified} - \text{introjected} - 2 \times \text{external}$.

At the end of the SLSQ, experimental students were also asked a series of open-ended questions about their experiences with ePEARL. These questions included items such as, "I like using ePEARL in my class because. . ." and "I did not like using ePEARL in my class because. . ." as well as "What I liked most about using ePEARL is. . ." and "What I liked least about ePEARL is. . ." Responses were coded to measure student enthusiasm as 1 = *Low*, 2 = *Medium*, 3 = *High*. A sample high-enthusiasm statement was, "It was a really fun thing to go on. I liked it so much." A sample medium-enthusiasm statement was, "It was an OK program to use to help us learn. I'm not a huge fan." A sample low-enthusiasm statement was, "I didn't like anything. It was annoying, and I'm so glad I don't have to go back on it." Enthusiasm statements were coded by two independent raters, and the four instances of disagreements were resolved through discussion.

At two points during the year (April and June), teachers completed an Implementation Fidelity Questionnaire (IFQ; Meyer et al., 2010, 2011) that asked them to report on how many hours a month they had been using ePEARL as well as describe what was going well and what were challenges they were facing. In order to measure implementation fidelity, Meyer et al.'s (2010, 2011) Implementation Assessment Protocol (IAP v2) was slightly revised. This protocol assessed the data reported on the IFQ and a sample of student portfolios in each classroom to determine the following: average number of artifacts, date range of use, and the degree to which students were using all of the available features of the software such as goal setting, attaching artifacts, feedback, and reflection. The IAP v2 allows each experimental classroom to be assigned a degree of implementation: low, medium, or high. For example, low-implementation classrooms would be those that reported less than 4 hr of ePEARL use each month; the student portfolios had zero to three student artifacts; and the artifacts would often be incomplete (i.e., work would be stored, but few of the SRL features would be used). Medium-high implementation

classrooms would be those that would report using ePEARL 5 or more hr each month; the student portfolio would have at least four artifacts; and the artifacts would include goals, content, and a reflection. Table 1 provides an overview of the IAP criteria for ePEARL implementation.

A research team member visited classrooms one to three times during the school year. These classroom visits served both data collection and implementation support purposes. The research team member presented a mini lesson on a feature of SRL and provided any necessary technological training and support. The content of these visits was documented in field notes that tracked how ePEARL was being used by the teachers and the students. At the end of the school year, teacher exit interviews were conducted in experimental classes using the Teacher Exit Interview Protocol (Meyer et al., 2010, 2011). This semistructured interview protocol was designed to explore the reasons for teachers' varying degrees and types of implementation, including their expectations, access to technology, support from administration and tech personnel, familiarity with portfolio pedagogy, knowledge of SRL processes, and time management issues.

Analyses

Two raters analyzed a random sample of each classroom's portfolios and independently assigned each classroom a rating of low, medium, or high implementation on the basis of the criteria outlined in the IAP v2. Raters achieved 90% agreement and together determined the IAP scores for each of the experimental teachers. All experimental classrooms were identified as medium- or high-implementation classrooms.

Cronbach's alpha, which measures internal consistency, was found to be .86 for both the SLSQ total pretest and posttest scores; alpha for the SRQ pretest was .91, and for the SRQ posttest, alpha was .88. Reliability for the six subscales of the SLSQ ranged from .81 to .88, whereas those of the four subscales for the SRQ ranged from .86 to .91. All effect size (ES) calculations follow Cohen's f (Cohen, 1988) for F ratios produced in analyses of variances,

$f = \sqrt{\frac{n^2}{1-n^2}}$. According to Cohen (1988), an f value of .10 can be considered as a small effect, .25 would be a medium effect, and .40 a large effect; however, these labels must be interpreted in the context of the research study; for example, a medium ES for a quasi-experimental study might be interpreted as more valuable than a medium ES for an experimental study. We report ESs for both multivariate as well as univariate analyses because the former may be misleading in that it is adjusted for covariance among outcome variables.

We used both design and statistical adjustments while inspecting and screening our data. From the initial experimental sample of 206 students, we decided to hold aside 39 (18.93%) students from the experimental group on the basis of their low enthusiasm for using the software because these students' achievement and self-regulation scores were significantly lower than those with either medium ($n = 84$ or 40.78%) or high enthusiasm ($n = 83$ or 40.29%). These low-enthusiasm students merit a separate analysis, which is detailed below, but are excluded from the main analyses as they were resistant to participating in the study and hesitant about using EPs for learning. Multivariate analyses using the pretest scores for CAT-4, the SLSQ, and SRQ as covariates; the

Table 1
Implementation Assessment Protocol

Criterion	Low	Medium	High
IFQ— hr/month	Hr ≤ 4	5–12	13 \leq hr
Avg. # artifacts	Artifacts ≤ 3	4–6	7 \leq artifacts
Date range of use	Entries span less than 60 days	Entries span 61–120 days	Entries span 121 days or more
Uses of EP's			
Planning: Goals & Strategies	<ul style="list-style-type: none"> ▪ 1 or no General Goals ▪ 1 or no Task Goals ▪ 1 or no Strategies 	<ul style="list-style-type: none"> ▪ At least 2 General Goals ▪ At least 3 artifacts have goals/strategies, content, & reflection • Goals & strategies may be vague, inappropriate, or may be attached to a grade/mark • Artifacts may be in only one subject area 	<ul style="list-style-type: none"> ▪ 3 or more General Goals—some may have been revised ▪ 4 or more artifacts have goals & strategies ▪ Goals & strategies are clearly defined and appropriate to the task
Doing: Content	<ul style="list-style-type: none"> ▪ Storage only • Incomplete entries 	<ul style="list-style-type: none"> • At least 3 artifacts have content ▪ Content is missing in some artifacts 	<ul style="list-style-type: none"> • Creative use of EP (different attachments, well-developed home page) ▪ 4 or more artifacts have content included (attachments, text editor, audio files) ▪ Artifacts included from multiple subject areas ▪ Multiple versions of artifacts ▪ 4 or more artifacts have reflections ▪ Reflections show deep thought about learning process and/or addresses goals & strategies
Reflecting	1 or no reflections	<ul style="list-style-type: none"> ▪ At least 3 artifacts have reflections ▪ Reflections are brief and generally vague (“I liked it, I had fun”) 	<ul style="list-style-type: none"> ▪ Teacher feedback on 4 or more artifacts • Feedback offers suggestions, asks questions, stimulates reflection • Feedback from peers includes constructive suggestions ▪ At least 3 items with reflection/selection explanation included
Feedback	No feedback	<ul style="list-style-type: none"> • Teacher feedback in fewer than 3 artifacts ▪ Feedback is summative only (gives a score/directive) ▪ Feedback from peers has a lot of “chat” 	
Presentation folder	Empty	<ul style="list-style-type: none"> • 1–2 items ▪ No reflection/selection reasons given or icons only selected 	

Note. IFQ = Implementation Fidelity Questionnaire; Avg. = average; EP = electronic portfolio.

posttest scores for the CAT-4, SLSQ, and SRQ as dependent variables; and the three levels of enthusiasm yielded a significant value of Pillai's trace of 1.116, $F(9, 588) = 38.70$, $p < .001$, $ES = .76$. Subsequent univariate tests showed that low-enthusiasm students ($M = 47.85$, $SD = 10.59$) scored significantly lower than their counterparts in the medium- ($M = 55.90$, $SD = 8.90$) and high-enthusiasm groups ($M = 57.40$, $SD = 10.57$) on their SLSQ posttest scores, $F(2, 196) = 3.77$, $p < .05$, $ES = .18$. Similarly, students in the low-enthusiasm group ($M = 7.76$, $SD = 2.41$) scored significantly lower than those in the medium-enthusiasm group ($M = 8.53$, $SD = 2.22$) on the CAT-4 achievement posttest score, $F(2, 200) = 3.77$, $p < .05$, $ES = .18$.

An additional reduction in sample size of 46 students occurred after we conducted a missing value analysis using SPSS software and excluded those who did not respond to items in either the CAT-4 achievement tests or the SRQ and SLSQ instruments. Analyses using SPSS version 19 revealed the values were missing at random, and hence, although removing these data might affect the power of the ensuing analyses, any parameter estimates would be unbiased. Finally, 79 control group participants were eliminated from the sample to establish pretest equivalence between the groups. After removal, results of a between-groups multivariate analysis of variance using pretest CAT-4, SRQ, and SLSQ scores

as dependent measures revealed a nonsignificant Pillai's trace of .002 ($p = .90$). This yielded a final sample of 319 (n for experimental group = 154, n for control group = 165) with no missing data for any variables.

All survey data were entered into SPSS version 19 by two graduate research assistants and verified for accuracy. For all measures, analyses were run using a multivariate analysis of covariance (MANCOVA) design, with the treatment (experimental, control) as the independent variable, pretest scores as covariates, and posttest scores as dependent variables. For all analyses, results of evaluation of assumptions of normality, homogeneity of variance-covariance matrices, linearity, and multicollinearity were satisfactory. Customized models were tested in the MANCOVA to ensure that the homogeneity of regression slopes assumption was met. There were no univariate or multivariate within-cell outliers at $\alpha = .001$. Questionnaire data were analyzed by item and aggregated to their respective scales, to obtain a fine-grained analysis of specific changes in self-regulation that occurred as a result of ePEARL use. These quantitative data, combined with an analysis of carefully selected student portfolios and classroom observations, provide a rich picture of how the use of ePEARL supports the academic achievement and self-regulation of learners who participated in the study.

Results

The first MANCOVA was conducted with the three posttest composite scores of the CAT-4, with SLSQ and SRQ treated as dependent variables, the EP-control treatment as the fixed factor, and the three pretest composite scores of the CAT-4, SLSQ, and SRQ as covariates. The multivariate test for main effect of the treatment on the dependent posttest measures was significant (Pillai's trace = .24), $F(6, 631) = 85.58$, $p < .001$, $ES = .35$. Follow-up tests showed that the experimental group significantly outperformed the control group on the SRQ posttest measures, $F(1, 314) = 10.09$, $p < .01$, $ES = .27$, after accounting for differences in pretest scores. See Table 2 for descriptive statistics, covariate values, and adjusted means for the pretest and posttest CAT-4, SRQ, and SLSQ composite measures. In addition, all the reported analyses were unchanged when the Huber-White correction (Huber, 1967; White, 1982) was applied to control for error due to nonnormal distributions of residuals in the dependent variables, with teacher, school, and province entered as clusters in the model.

A second MANCOVA was conducted with the CAT-4 posttest subscales of ideas presented in students' response to text, support for response to text, content presented in the writing assignment, as well as content management in writing skills as dependent variables, treatment versus control groups as fixed factors, and the CAT-4 pretest subscales as covariates. Neither the SLSQ nor the SRQ measures were included in these analyses. The multivariate test for the main effect of treatment on the CAT-4 posttest subscales after having accounted for differences in pretest subscale scores was significant (Pillai's trace = .144), $F(8, 624) = 6.33$, $p < .001$, $ES = .27$. Post hoc tests revealed that students using ePEARL had significantly greater posttest scores compared with controls in the subscales of providing support for response to text, $F(1, 313) = 11.79$, $p < .01$, $ES = .27$; content presented in the

Table 3

Descriptive Statistics and Adjusted Means From MANCOVA With CAT-4 Subscale Measures for Experimental and Control Groups

CAT-4 subscale means	Experimental group		Control group	
	Pre ^a	Post ^b	Pre ^a	Post ^b
Response to Text: Ideas	1.71	1.99 (1.98)	1.73	2.04 (2.03)
SD	.61	.66	.53	.62
Response to Text: Support	1.71	2.10 (2.01)	1.68	1.83 (1.86)
SD	.68	.70	.59	.74
Writing: Content	1.72	2.10 (2.11)	1.75	1.88 (1.88)
SD	.67	.70	.62	.63
Writing: Content Management	1.64	2.00 (1.95)	1.63	1.73 (1.71)
SD	.67	.69	.57	.64

Note. $n = 154$ (experimental), 165 (control). MANCOVA = multivariate analysis of covariance; CAT-4 = Canadian Achievement Test, fourth edition.

^a Pretest covariates in the MANCOVA were set at the sample mean. ^b Adjusted means for posttest scores calculated in MANCOVA appear in parentheses.

writing assignment, $F(1, 313) = 9.78$, $p < .01$, $ES = .25$; and content management in writing skills, $F(1, 313) = 12.00$, $p < .01$, $ES = .27$. See Table 3 for detailed descriptive statistics, covariate values, and adjusted means for the CAT-4 subscale responses.

A third MANCOVA was conducted with the SLSQ posttest subscales as dependent variables, experimental versus control group as the fixed factor, and SLSQ pretest subscales as covariates, and these yielded a significant Pillai's trace of .20, $F(12, 612) = 33.44$, $p < .001$, $ES = .33$. Follow-up tests revealed that participants using ePEARL showed significantly higher posttest scores than the controls on five of six subscales of the SLSQ: Goal Setting, $F(1, 311) = 11.58$, $p < .01$, $ES = .20$; Strategy Planning,

Table 2

Descriptive Statistics and Adjusted Means From MANCOVA With Achievement and Self-Regulation Composite Measures for Experimental and Control Groups

Composite score means	Experimental group		Control group	
	Pre ^a	Post ^b	Pre ^a	Post ^b
CAT-4	6.79	8.19 (8.09)	6.79	7.48 (7.37)
SD	2.06	2.23	1.46	1.96
SRQ ^c	-22.89	-17.94 (-20.59)	-22.41	-29.83 (-25.45)
SD	17.23	17.24	18.19	19.00
SLSQ	56.78	57.80 (57.45)	56.57	55.46 (55.83)
SD	10.34	9.52	10.64	10.63

Note. $n = 154$ (experimental), 165 (control). MANCOVA = multivariate analysis of covariance; CAT-4 = Canadian Achievement Test, fourth edition; SRQ = Academic Self-Regulation Questionnaire; SLSQ = Student Learning Strategies Questionnaire.

^a Pretest covariates in the MANCOVA were set at the sample mean. ^b Adjusted means for posttest scores calculated in MANCOVA appear in parentheses. ^c The SRQ Relative Autonomy Index = $2 \times$ intrinsic + identified - introjected - $2 \times$ extrinsic; this means that Intrinsic and Identified subscales contribute positively, whereas the Introjected and Extrinsic subscales reduce the self-regulation score in learners. The less negative the composite SRQ score, the better the overall relative autonomy index of the learner.

Table 4

Descriptive Statistics and Adjusted Means From MANCOVA With SLSQ Subscale Measures for Experimental and Control Groups

SLSQ subscale means	Experimental group		Control group	
	Pre ^a	Post ^b	Pre ^a	Post ^b
Goal Setting	7.04	7.60 (7.49)	6.79	6.83 (6.75)
SD	1.63	1.50	1.75	1.75
Strategy Planning	7.95	8.10 (7.95)	8.01	7.77 (7.72)
SD	1.59	1.46	1.34	1.35
Self-Observation	14.77	14.99 (14.82)	14.90	14.48 (14.35)
SD	2.52	2.51	2.55	2.82
Self-Instruction	11.11	10.81 (10.74)	10.86	10.41 (10.26)
SD	2.17	2.00	2.36	2.50
Feedback from Adults	8.35	8.33 (8.34)	8.50	8.43 (8.41)
SD	1.64	1.72	1.41	1.50
Self-Evaluation	7.56	7.97 (7.79)	7.51	7.54 (7.60)
SD	1.50	1.45	1.63	1.56

Note. $n = 154$ (experimental), 165 (control). MANCOVA = multivariate analysis of covariance; SLSQ = Student Learning Strategies Questionnaire.

^a Pretest covariates in the MANCOVA were set at the sample mean. ^b Adjusted means for posttest scores calculated in MANCOVA appear in parentheses.

Table 5
Descriptive Statistics and Adjusted Means From MANCOVA With SRQ Subscale Measures for Experimental and Control Groups

SRQ subscale means ^c	Experimental group		Control group	
	Pre ^a	Post ^b	Pre ^a	Post ^b
Extrinsic	27.70	24.84 (25.75)	27.31	26.95 (26.77)
SD	5.73	5.37	5.10	5.19
Introjected	27.41	24.62 (24.59)	26.30	27.00 (26.01)
SD	6.11	5.45	5.77	5.31
Identified	23.30	23.98 (22.25)	23.33	20.17 (21.01)
SD	3.97	3.74	4.22	4.05
Intrinsic	18.31	20.04 (18.20)	17.59	15.45 (15.46)
SD	3.95	5.23	5.46	5.31

Note. $n = 154$ (experimental), 165 (control). MANCOVA = multivariate analysis of covariance; SRQ = Academic Self-Regulation Questionnaire.

^a Pretest covariates in the MANCOVA were set at the sample mean. ^b Adjusted means for posttest scores calculated in MANCOVA appear in parentheses. ^c The SRQ Relative Autonomy Index = $2 \times \text{intrinsic} + \text{identified} - \text{introjected} - 2 \times \text{extrinsic}$; this means that Intrinsic and Identified subscales contribute positively, whereas the Introjected and Extrinsic subscales reduce the self-regulation score in learners. The lesser the extrinsic and introjected scores, the more autonomy the learner possesses.

$F(1, 311) = 10.55, p < .01, ES = .18$; Self-Observation, $F(1, 311) = 6.33, p < .01, ES = .15$; Self-Instruction, $F(1, 311) = 5.33, p < .01, ES = .14$; and Self-Evaluation, $F(1, 311) = 5.32, p < .01, ES = .14$. See Table 4 for the descriptive statistics, covariate values, and adjusted means for the SLSQ subscales.

A fourth MANCOVA of the SRQ subscales as dependent measures, experimental versus control group as the fixed factor, and the SRQ pretest subscales as covariates yielded a significant Pillai's trace of .20, $F(8, 624) = 8.54, p < .001, ES = .32$. Post hoc tests showed that students who used the EP had significantly better posttest scores than the controls for all four SRQ subscales, namely, Extrinsic, $F(1, 313) = 11.33, p < .01, ES = .26$; Introjected, $F(1, 313) = 11.66, p < .01, ES = .26$; Identified, $F(1, 313) = 13.10, p < .01, ES = .28$; and Intrinsic, $F(1, 313) = 12.54, p < .01, ES = .27$. See Table 5 for the descriptive statistics, covariate values, and adjusted means for these SRQ subscales.

Illustration of ePEARL Use

A review of the student work stored in ePEARL using the IAP v2 demonstrated that the teachers used ePEARL to help students develop SRL skills as well as literacy, information and communication technology (ICT), and other content area skills. Some sample projects included a reader's theatre, writing a fable, designing simple machines, building an insulating machine, and constructing a model of a First Nations village. For each artifact, students were prompted to describe the task, set task goals, identify strategies to accomplish these goals, as well as reflect on their progress toward these goals. The following excerpt is a sample of one student's artifact titled "Simple machines."

Criteria

- 1: I will use all 5 simple machines in the playground.
- 2: I will be creative and use the simple machines to create new ideas.
- 3: My inventions will be detailed and they will work.
- 4: The playground will be sturdy and neatly assembled.
- 5: I will be able to explain how each simple machine works in my playground.

Goals

Task Goals Updated 01/23/09

I will use materials that actually work the way the simple machine should.

Strategies Updated 01/23/09.

- 1: I will use my notes as a resource for reminders.
- 2: I will use my blueprint as a guide.
- 3: I will share ideas with my partner.
- 4: I will test materials before I put them in my playground.
- 5: I will test all my machines before I share my playground with the class.

Content

[images of simple machines]

Reflections Updated 02/10/09

Me and Joseph did so well, I've got my goal done by due date. We used all 5 simple machines it was sturdy. I'm pretty sure my science grade will be great. Next time I need to get Joseph to stop fooling around and help me, it's getting annoying.

Teacher Feedback

It's hard to work with someone who isn't putting in the same effort. Would you choose a different partner next time? You met your criteria and met your goal at the same time! Way to go!

This text shows how the students used ePEARL to document their plan and monitor their progress as they worked on their project with a partner. The teacher feedback at the end gives evidence of how the teachers worked to give specific feedback to guide their reflections and to prompt additional learning on the basis of the results of this lesson.

While visiting the classrooms, the researchers provided sample lessons and documented examples of how the teachers were supporting student development of SRL skills. The following excerpt provides a sample of how one teacher introduced the topic of reflection.

[The teacher] presented a lesson on the SMART board about portfolios and the purposes of them. She had created all kinds of fun games in the SMART board that had students engaged and participating in identifying the differences between a portfolio and a random collection of work/notes/assignments: "shows learning" "you reflect" "you select." She then pulled up a few sample artifacts and showed strong examples of student work that had good task goals, strategies, criteria,

Task Description

We will use the simple machines of pulleys, rollers, levers, wheels and axles, and incline plane. We will build the model using lots of different materials in a small cardboard box.

and reflections. She talked about the importance of reading the goals and reflections for her as a teacher then showed the ePEARL video: reflecting (works in progress) and talked about how she reflects daily and how she asks the students to reflect on their work.

She then showed the rubric that she had created for the “personal heroes project” and had sent to all the students via the portal. She explained to them how to attach the rubric and said that was a requirement for this assignment, as it was a new ICT skill they needed to practice. She then introduced another SMART board game to talk about strategies (word scramble to answer a question: “What helps us work to achieve our goals?”). She explained that the focus for this assignment was to learn to attach the rubric and to reflect on their progress as they worked on this project. She told them that this was a social studies artifact [indicating the folder they should assign it to] and told them to choose a partner to work with on the computer. They should help each other enter their plan (they had already completed planning sheets). When they were both finished, they were instructed to work on their graphic organizer for the project (paper and pencil). Students had about 20 minutes to work, and most got the planning done (2009–02-22- field notes).

This lesson demonstrated that the teacher was using the instructional supports provided via the software (instructional videos and planning sheets) to provide focused instruction on particular SRL skills. The teacher interview data indicated that the teachers found these visits to be very helpful, and the supports designed to support the software provided them the tools and information necessary to provide effective instruction on the SRL process.

Discussion

The results of this study provide important confirmatory evidence of the positive impact of the classroom use of EPs on students’ literacy skills and SRL strategies. Enthusiastic students who used ePEARL in medium- or high-implementation classrooms demonstrated moderate-size learning gains on a standardized literacy measure and reported positive changes in key SRL skills. Whereas other research explores the processes involved in the development and use of SRL, this study adds an important element by providing convincing evidence that a theoretically based knowledge tool, when wisely and well implemented by classroom teachers, can have a meaningful impact on learning.

EPs are promoted as knowledge tools that are designed to facilitate the integration of technology in classrooms by being fully embedded into classroom life rather than merely added to it. In contrast to the longitudinal study by Meyer et al. (2010), in the current investigation medium or high implementation of the EP was achieved in all the experimental classrooms. In these classrooms, the positive impact on learning and self-regulation that a process EP can achieve was documented, despite using an instrument—the constructed response subtest of the CAT-4—that is not especially sensitive to small, subtle changes in student literacy skills. At the same time, it was noted that not all students were uniformly enthusiastic about ePEARL. Fortunately, the percentage of low-enthusiasm students was much smaller (less than 20%) than the percentage of medium- and high-enthusiasm students (about 80%), attesting to the applicability and acceptance of EPs among the majority of students, at least through elementary school. The low-enthusiasm students did not show the same academic improvements or benefits from the standpoint of developing self-

regulatory competencies as their counterparts, as our statistical analyses point out, and merit further investigation in future research, especially given that motivation is a key component of self-regulation.

Prior research and the reviews of research on teaching students to self-regulate (Dignath & Buettner, 2008; Dignath et al., 2008) point to both the benefits of student self-regulation and instructional interventions that enhance these skills in students. The current quasi-experiment documents the feasibility of developing a knowledge tool that can be used with fidelity by classroom teachers and received with enthusiasm by the majority of students. Teachers, and not researchers, successfully implemented an SRL instructional program with a significant impact on students’ SRL processes, leading to changes in their reading and writing skills, although mathematics might have been an easier subject area to affect change. This research provides good evidence with regards to (a) ameliorating concerns (e.g., Barrett, 2007) about the lack of evidence of the impact of EPs on student learning and (b) reducing concerns (e.g., Zimmerman, 2008) about the lack of evidence of the impact of SRL knowledge tools on student learning.

Future Directions

Winne, Hadwin, and Gress (2010) discuss the importance of socially shared self-regulation and coregulation, emphasizing that knowledge building often occurs in collaboration with others. ePEARL allows students to collaborate with others, but the tool does not yet scaffold the strategies for joint productivity as well as it might. Revisions to the software to encourage use within cooperative and collaborative learning environments are necessary and would fit seamlessly within student-centered contexts.

Second, further research should explore the extended use of ePEARL by students and teachers. Also of value would be a follow-up investigation concerning whether learning gains and student SRL changes promoted by ePEARL use during one school year were sustained over time.

Third, ePEARL has been linked by the authors with other evidence-based educational software, including an early literacy tool (ABRACADABRA), an inquiry and information literacy tool (ISIS-21), and the prototype of an early mathematics tool (ELM/ORME). Ongoing research will document the challenges and complexities of using two tools simultaneously, one to scaffold SRL strategies and another to scaffold the learning of curricular content, and whether this can be done together both effectively and efficiently.

Fourth, Idan, Abrami, Wade, and Meyer (2011) developed ePEARL for adult learners (i.e., senior secondary, vocational, and postsecondary teachers and students) that builds on the design of the previous three levels of ePEARL by explicitly scaffolding detailed aspects of the motivational, cognitive, and metacognitive aspects of SRL. Initial concerns regarding ePEARL Level 4 focus on usability, including the acceptance of a more complex and comprehensive tool by teachers and students.

Lastly, ePEARL has been adapted for use by studio music teachers and their students, called iSCORE, and research is underway on its use within arts education (Upitis, Abrami, Brook, Troop, & Varela, 2012).

Cautions and Limitations

The strengths of this research include the size and geographic diversity of the participants; the successful integration of the tool as part of classroom practice in medium- and high-implementation classrooms; the length of the study; and the use of a standardized achievement measure compatible with the underlying philosophy of portfolios. The weaknesses of this research relate mostly to research design and instrumentation. We recognize that the self-report nature of SRL measures have been shown to be somewhat inaccurate representations of actual SRL (Jamieson-Noel & Winne, 2003), and in our future research, we hope to collect more behavioral or log-file data from students' use of software to better understand SRL competencies (Shaikh et al., 2012; Venkatesh & Shaikh, 2008).

A strong quasi-experimental design was used but not a true experimental design. Homogeneity of covariance matrices and homogeneity of regression assumptions for all the analyses of covariance were verified. Although it might be argued that the study's ecological validity was affected due to the suppression of data from the ePEARL group students with low enthusiasm for using the software, our analysis demonstrates that they significantly underperformed both on the achievement and self-regulation measures as compared with the students with medium or high enthusiasm, and so their data were best analyzed separately.

Implementation Issues in the Use of EPs

On the basis of studying ePEARL use in classrooms both in this study and over several years (Abrami & Barrett, 2005; Abrami et al., 2006; Abrami, Wade, et al. 2008; Bures, Barclay, Abrami, & Meyer, 2009; Meyer, Abrami, & Wade, 2009; Meyer et al., 2010, 2011; Wade et al., 2005), some valuable lessons were learned:

1. The use of portfolios should be a school-based or board-(district-) based initiative and integrated into regular classroom teaching. Use of the EP in one or two classrooms once or twice a week will have a smaller impact.

2. The use of portfolios should begin early in students' educational experience and not be short lived. The processes of self-regulation and approaches to pedagogy that electronic portfolios support require time for younger students to learn and effort for older students to make the transition from traditional, teacher-directed methods.

3. The regular and systematic use of EPs should be undertaken when students work on novel, complex, and challenging tasks. Unimportant and simple tasks do not require a knowledge tool that provides the degree of learner scaffolding apparent in EP software.

4. Teachers need to develop facility with portfolio processes, and they should be supported with appropriate professional development and administrative support.

5. EPs provide the means to scaffold teachers and students in the portfolio process and better encourage self-regulation, although these tools are not a sufficient condition for change.

6. Students and teachers must believe that the change to using a process portfolio is valued and necessary for authentic, more meaningful learning. The "will" component of SRL is as important as the "skill" component.

This last observation is perhaps the most important. This research project has operated under the belief that learners benefit

from knowledge tools for learning and that they need to learn how to use them in order to experience achievement gains. However, do learners and their teachers see the value in using these tools for learning? And do learners want to learn how to use them? These questions touch on a number of dilemmas in contemporary education—the challenges of creating and sustaining effective student-centered learning environments, the difficulties in integrating technology in classrooms, and the obstacles to switching pedagogy from emphasizing *what* content is to be learned to emphasizing *how* content is to be learned. More particularly, answering these questions may help explain the failure of other researchers to document wide-scale and faithful implementations of other EPs and the inability to document the impact of tool use on teachers and their students.

It is hoped that the findings of this research will encourage school leaders and teacher educators to recognize the value and importance of EPs to support SRL. This study indicates that students improve in their writing and certain SRL skills when an EP is used regularly and appropriately throughout the school year. In order to encourage the effective integration of EPs, or for other technological and pedagogical innovations to happen widely and well, school leaders, teacher educators, and pedagogical support staff need to provide consistent positive support to teachers as they learn to teach with new technologies and work within the changing realities of their school environments.

There have been many calls to transform classrooms to become student centered and to use technology and knowledge tools as a means to promote active, engaged, and reflective learning by students. This research has provided some evidence that EPs in general, and ePEARL specifically, answer these calls.

References

- Abrami, P. C. (2010). On the nature of support in computer supported collaborative learning using gstudy. *Computers in Human Behavior*, 26, 835–839. doi:10.1016/j.chb.2009.04.007
- Abrami, P. C., & Aslan, O. (2007). *The student learning strategies questionnaire (SLSQ)*. Unpublished instrument. Retrieved from http://doe.concordia.ca/cslp/cslp_cms/?q=node/49
- Abrami, P. C., & Barrett, H. (2005). Directions for research and development on electronic portfolios. *Canadian Journal of Learning and Technology*, 31, 1–15. Retrieved from <http://www.cjlt.ca/index.php/cjlt/article/viewArticle/92/86>
- Abrami, P. C., Bernard, R. M., Bures, E. M., Borokhovski, E., & Tamim, R. (2011). Interaction in distance education and online learning: Using evidence and theory to improve practice. *Journal of Computing in Higher Education*, 23, 82–103. doi:10.1007/s12528-011-9043-x
- Abrami, P. C., Bernard, R. M., Wade, A., Schmid, R. F., Borokhovski, E., Tamim, R., & Peretiatkiewicz, A. (2006). A review of e-learning in Canada: A rough sketch of the evidence, gaps and promising directions. *Canadian Journal of Learning and Technology*, 32, 1–70. Retrieved from <http://www.cjlt.ca/index.php/cjlt/article/view/27/25>
- Abrami, P. C., Savage, R. S., Wade, A., & Higgs, G. (2008). Using technology to assist children learning to read and write. In T. Willoughby & E. Wood (Eds.), *Children's learning in a digital world* (pp. 129–172). Oxford, UK: Blackwell Publishing.
- Abrami, P. C., Wade, A., Pillay, V., Aslan, O., Bures, E., & Bentley, C. (2008). Encouraging self-regulated learning through electronic portfolios. *Canadian Journal of Learning and Technology*, 34, 93–117. Retrieved from <http://www.cjlt.ca/index.php/cjlt/article/view/507/238>

- Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice*, 11, 36–44. doi:10.1111/j.1745-3992.1992.tb00230.x
- Avramidou, L., & Zembal-Saul, C. (2003). Exploring the influence of web-based portfolio development on learning to teach elementary science. *Journal of Technology and Teacher Education*, 11, 415–442. Retrieved from <http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED467271>
- Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist*, 40, 199–209. doi:10.1207/s15326985ep4004_2
- Barrett, H. C. (2007). Researching electronic portfolios and learner engagement: The REFLECT initiative. *Journal of Adolescent & Adult Literacy*, 50, 436–449. doi:10.1598/JAAL.50.6.2
- Barrett, H. C. (2008, July). *The REFLECT initiative: A research project to assess the impact of electronic portfolios on student learning, motivation and engagement in secondary schools (Final report presented to National Educational Computing conference)*. Retrieved from <http://electronicportfolios.com/portfolios.html>
- Barrett, H. C. (2009). Online personal learning environments: Structuring electronic portfolios for lifelong and life wide learning. *On the Horizon*, 17, 142–152. Retrieved from http://docs.google.com/Doc?id=dd76m5s2_39fsmjddk
- Bernard, R. M., Bethel, E. C., Abrami, C., & Wade, A. (2007). Introducing laptops to children: An examination of ubiquitous computing in grade three reading, language, and mathematics. *Canadian Journal of Learning and Technology*, 33, 49–74.
- Bures, E. M., Barclay, A., Abrami, P. C., & Meyer, E. (2009, August). *Contextualizing student assessment: How can teachers effectively assess electronic portfolios?* Paper presented at the European Association of Research and Learning in Instruction (EARLI) conference, Amsterdam, the Netherlands.
- Campuzano, L., Dynarski, M., Agodini, R., Rall, K., & Pendleton, A. (2009). Effectiveness of reading and mathematics software products: Findings from two student cohorts (Report). Washington, DC: U.S. Department of Education, Institute of Education Sciences.
- Canadian Council on Learning. (2008). *E-Learning in Canada: The promise and potential (Thematic report)*. Ottawa: The Council.
- Canadian Test Centre. (2008). *Canadian Achievement Tests, fourth edition (CAT - 4)*. Markham, ON: Author.
- Carney, J. (2005). *What kind of electronic portfolio research do we need?* Paper presented at the Society for Information Technology and Teacher Education (SITE) conference, Phoenix, AZ. Retrieved from <http://it.wcu.edu/carney/Presentations/SITE05/ResearchWeNeed.pdf>
- CEO Forum on Education and Technology. (2001). *Key building blocks for student achievement in the 21st century: Assessment, alignment, accountability, access, analysis*. Retrieved from <http://www.ceoforum.org/downloads/forum3.pdf>
- Chung, M.-K. (2000). The development of self-regulated learning. *Asia Pacific Education Review*, 1, 55–66. doi:10.1007/BF03026146
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Coulombe, S., Tremblay, J. F., & Marchand, S. (2004). *Literacy scores, human capital and growth across fourteen OECD countries* (Catalogue No. 89–552–MIE, no. 11). Ottawa: Statistics Canada.
- Cuban, L. (1993). Computers meet classroom: Classroom wins. *Teachers College Record*, 95, 185–210.
- Cuban, L., Kirkpatrick, H., & Peck, C. (2001). High access and low use of technologies in high school classrooms: Explaining an apparent paradox. *American Educational Research Journal*, 38, 813–834. doi:10.3102/00028312038004813
- Dignath, C., & Buettner, G. (2008). Components of fostering self-regulated learning among students: A meta-analysis of intervention studies at primary and secondary level. *Metacognition and Learning*, 3, 231–264. doi:10.1007/s11409-008-9029-x
- Dignath, C., Buettner, G., & Langfeldt, H. P. (2008). How can primary school students learn self-regulated learning strategies most effectively? A meta-analysis of self-regulation training programmes. *Educational Research Review*, 3, 101–129. doi:10.1016/j.edurev.2008.02.003
- Dynarski, M., Agodini, R., Heaviside, S., Nowak, T., Carey, N., Campuzano, L., . . . Sussex, W. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort (Report)*. Washington, DC: U. S. Department of Education, Institute of Education Sciences.
- European Union Council. (2002, June). Council resolution of 27 June 2002 on lifelong learning (no. 32002G0709(01)). Retrieved from EUR-Lex website: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002G0709%2801%29:EN:NOT>
- Hillyer, J., & Lye, T. C. (1996). Portfolios and second graders' self-assessments of their development as writers. *Reading Improvement*, 33, 148–159.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (pp. 221–233). Berkeley: University of California Press.
- Idan, E., Abrami, P. C., Wade, A., & Meyer, E. (2011, March). *Designing for the development of self-regulation: A web-based electronic portfolio for adult learners*. Paper presented at the International Technology, Education and Development conference, Valencia, Spain.
- Jamieson-Noel, D. L., & Winne, P. H. (2003). Comparing self-reports to traces of studying behavior as representations of students' studying and achievement. *German Journal of Educational Psychology*, 17, 159–171.
- Knighton, T., Brochu, P., & Gluszynski, T. (2010). *Measuring up: Canadian results of the OECD PISA study - The performance of Canada's youth in reading, mathematics and science; 2009 first results for Canadians aged 15* (Catalogue No. 81–590-X). Ottawa, Ontario: Human Resources and Skills Development Canada, Council of Ministers of Education, Canada and Statistics Canada.
- MacIsaac, D., & Jackson, L. (1994). Assessment processes and outcomes: Portfolio construction. *New Directions for Adult and Continuing Education*, 62, 63–72. doi:10.1002/ace.36719946208
- Mayer, R. E. (2001). *Multimedia learning*. New York, NY: Cambridge University Press. doi:10.1017/CBO9781139164603
- Mayer, R. E. (2008). *Learning and instruction* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice-Hall.
- Meyer, E., Abrami, P. C., & Wade, A. (2009, April). *Electronic portfolios in the classroom: Factors impacting teacher's integration of new technologies and new pedagogies*. Paper presented at the annual meeting of the American Educational Research Association (AERA) San Diego, CA.
- Meyer, E., Abrami, P. C., Wade, A., Aslan, O., & Deault, L. (2010). Improving literacy and metacognition with electronic portfolios: Teaching and learning with ePEARL. *Computers & Education*, 55, 84–91. doi:10.1016/j.compedu.2009.12.005
- Meyer, E., Abrami, P. C., Wade, A., & Scherzer, R. (2011). Electronic portfolios in the classroom: Factors impacting teachers' integration of new technologies and new pedagogies. *Technology, Pedagogy and Education*, 20, 191–207. doi:10.1080/1475939X.2011.588415
- Organization for Economic Co-Operation and Development. (2010). *PISA 2009 Results: What students know and can do – Student performance in reading, mathematics and science (Volume I)*. Paris, France: OECD–UNESCO. doi: [10.1787/9789264091450-en](http://dx.doi.org/10.1787/9789264091450-en)
- Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist*, 36, 89–101. doi: 10.1207/S15326985EP3602_4
- PHP—Hypertext Preprocessor (Version 5.X) [Scripting language]. Retrieved from <http://www.php.net>

- Pintrich, P. R. (1999). The role of motivation in promoting and sustaining self-regulated learning. *International Journal of Educational Research*, 31, 459–470. doi:10.1016/S0883-0355(99)00015-4
- Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95, 667–686. doi:10.1037/0022-0663.95.4.667
- Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, 57, 749–761. doi:10.1037/0022-3514.57.5.749
- Shaikh, K., Zuberi, A., & Venkatesh, V. (2012). Exploring counter-theoretical instances of graduate learners' self-regulatory processes in online learning environments. *International Journal of Technologies in Higher Education*, 9, 6–19.
- Tobias, S., & Duffy, T. M. (Eds.). (2009). *Constructivist instruction: Success or failure?* New York, NY: Routledge Taylor & Francis Group.
- Ungerleider, C., & Burns, T. (2003, October). *A systematic review of the effectiveness and efficiency of networked ICT in education (A state of the field report to the Council of Ministers of Education, Canada and Industry Canada)*. Retrieved from <http://www.cmec.ca/stats/SystematicReview2003.en.pdf>
- Uptis, R., Abrami, P. C., Brook, J., Troop, M., & Varela, W. (2012, January). *Learning to play an instrument by enhancing self-regulatory skills with digital portfolio tools*. Paper presented at the tenth annual Hawaii International Conference on Arts & Humanities, Honolulu, HI.
- Venkatesh, V., & Shaikh, K. (2008). Investigating task understanding in online repositories equipped with topic map indexes: Implications for improving self-regulatory processes in graduate learners. *International Journal of Technologies in Higher Education*, 5, 22–35.
- Venkatesh, V., & Shaikh, K. (2011). Uncovering relationships between task understanding and monitoring proficiencies in post-secondary learners: Comparing work task and learner as statistical units of analyses. *Education Research International*, 2011, Article ID 735643, 11 pages. doi:10.1155/2011/735643
- Wade, A., Abrami, P. C., & Sclater, J. (2005). An electronic portfolio to support learning. *Canadian Journal of Learning and Technology*, 31, 33–50. Retrieved from <http://www.cjlt.ca/index.php/cjlt/issue/view/13>
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25. doi:10.2307/1912526
- Winne, P. H. (1995). Inherent details in self-regulated learning. *Educational Psychologist*, 30, 173–187. doi:10.1207/s15326985ep3004_2
- Winne, P. H., Hadwin, A. F., & Gress, C. L. Z. (2010). The Learning Kit project: Software tools for supporting and researching regulation of collaborative learning. *Computers in Human Behavior*, 26, 787–793. doi:10.1016/j.chb.2007.09.009
- Wozney, L., Venkatesh, V., & Abrami, P. C. (2006). Implementing computer technologies: Teachers' perceptions and practices. *Journal of Technology and Teacher Education*, 14, 173–207.
- Zeichner, K., & Wray, S. (2001). The teaching portfolio in US teacher education programs: What we know and what we need to know. *Teaching and Teacher Education*, 17, 613–621. doi:10.1016/S0742-051X(01)00017-8
- Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, 25, 3–17. doi:10.1207/s15326985ep2501_2
- Zimmerman, B. J. (1998). Developing self-fulfilling cycles of academic regulation: An analysis of exemplary instructional models. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulated learning: From teaching to self-reflective practice* (pp. 1–20). New York, NY: Guilford Press.
- Zimmerman, B. J. (2000). Attaining self-regulation: A social-cognitive perspective. In M. Boekaerts & P. R. Pintrich (Eds.), *Handbook of self-regulation* (pp. 13–39). New York, NY: Academic Press. doi:10.1016/B978-012109890-2/50031-7
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments and future prospects. *American Educational Research Journal*, 45, 166–183. doi:10.3102/0002831207312909
- Zimmerman, B. J., & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. *American Educational Research Journal*, 31, 845–862. doi:10.3102/00028312031004845
- Zimmerman, B. J., & Martinez-Pons, M. (1988). Construct validation of a strategy model of students self-regulated learning. *Journal of Educational Psychology*, 80, 284–290. doi:10.1037/0022-0663.80.3.284
- Zimmerman, B. J., & Schunk, D. H. (2011). Self-regulated learning and performance: An introduction and overview. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation and performance* (pp. 1–12). New York, NY: Routledge.
- Zimmerman, B. J., & Tsikalas, K. E. (2005). Can computer-based learning environments (CBLEs) be used as self-regulatory tools to enhance learning? *Educational Psychologist*, 40, 267–271. doi:10.1207/s15326985ep4004_8
- Zubizarreta, J. (2004). *The learning portfolio: Reflective practice for improving student learning*. Bolton, England: Anker.

(Appendices follow)

Appendix A**Jump Start!****6 lessons for getting started with ePEARL (Levels 2&3)****Purpose:**

To introduce teachers & students to ePEARL and using electronic portfolios in K-12 classrooms.

Objectives:

To help teachers effectively introduce the basic features of ePEARL and the self-regulated learning process to their students.

Time required: 6 lessons (45 – 60 minutes each)

Materials required:

- 1) Jump Start lesson plans
- 2) Internet connected computer, speakers & projector or SMART board
- 3) List of student usernames and passwords
- 4) Mobile lab or reserve the computer lab

Topics addressed:

- 1) Lesson 1: Introduction & Help
 - a. What are electronic portfolios?
 - b. How does ePEARL work?
 - c. How do I log in and personalize my ePEARL?
 - d. How do I get help if I'm stuck?
- 2) Lesson 2: General Goals & Help
 - a. What are General Goals?
 - b. How to set good General Goals
 - c. How to input General Goals in ePEARL
 - d. How to get help setting General Goals
- 3) Lesson 3: Organizing your ePEARL
 - a. Why do we need to get organized?
 - b. Using the agenda
 - c. How to organize your work into folders
 - d. How to manage colour codes

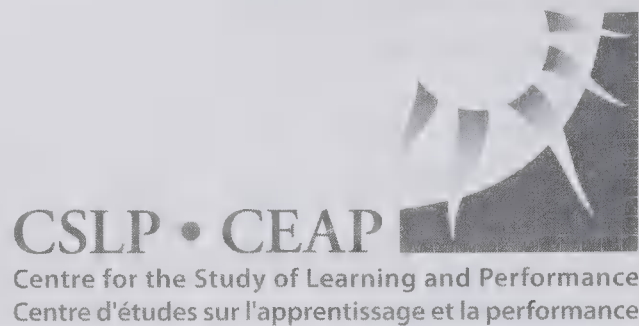
(Appendices continue)

- 4) Lesson 4: Planning - Starting a new artifact
 - a. Why is it important to plan your task?
 - b. What do I fill in here? understanding the terms:
 - i. Description
 - ii. Criteria & rubrics
 - iii. Task Goals
 - iv. Strategies
 - c. Helping students identify useful task goals and strategies
 - d. How to do this in ePEARL
- 5) Lesson 5: Doing
 - a. Composing in the text editor
 - b. Recording readings or music with the audio recorder
 - c. Attaching multimedia files
 - d. Reflecting on your task goals & strategies as you work
- 6) Lesson 6: Reflecting on works in progress and completed works
 - a. Sharing works with peers
 - b. Giving constructive feedback
 - c. Revising work based on feedback
 - d. Saving as a new version

Planning Tips:

1. **Train** 4-5 tech-savvy **students** on ePEARL the day before you introduce each lesson so they can assist you with any questions or challenges the students have.
2. **Set up** the equipment the afternoon **before** and practice logging in to make sure everything works properly, including your own username & password.
3. **Preview** the virtual tutorial chapters and instructional **videos** on each topic before presenting the lesson to the students. These will provide you with a general overview of the key software features and teaching ideas that will help you to present that lesson.
4. **Join** the **WIKI** – there are additional materials on this interactive forum to help with these lessons. You need to email a request to join this wiki and it may take 2-3 days to activate your account. To join write: emeyer@education.concordia.ca

(Appendices continue)

Appendix II

CENTRE FOR THE STUDY OF LEARNING AND PERFORMANCE
McConnell Building, 1455 de Maisonneuve Blvd. W., LB-581
Montreal, Quebec, Canada H3G 1M8
Tel: (514) 848-2424 x2020

Learning Strategies Questionnaire

This questionnaire is part of a study being conducted by the Centre for the Study of Learning and Performance at Concordia University in Montreal, Quebec. We would like to know more about how you are learning this year. This questionnaire will help us learn about the strategies you are using in your class to help you with your work.

Please answer the questions on the next page. **There is no right or wrong answer.** Your answers are confidential (no one that you know will be told what you answered). Your teacher will not have access to your answers. You have the right to refuse, to participate, or to withdraw (stop answering the questions) at any time. However, your experiences and opinions are important, and will help us understand teaching from your point of view.

Thank you for your collaboration!

Vanitha Pillay, Research Coordinator, CSLP
Phil Abrami, Professor and Director, CSLP

(Appendices continue)

PERSONAL INFORMATION

- Name: _____
- Gender: Boy _____ Girl _____
- School: _____ Grade _____

INSTRUCTIONS

Please circle the most appropriate response when answering the questions.
In my class...

1. I set my own learning goals (I decide what I need to learn).

Strongly Disagree Disagree Undecided Agree Strongly Agree

2. I set my own process goals (I list what I need to do to achieve my learning goals).

Strongly disagree Disagree Undecided Agree Strongly agree

3. I identify strategies for achieving my goals.

Strongly disagree Disagree Undecided Agree Strongly agree

4. I revise my goals when necessary.

Strongly disagree Disagree Undecided Agree Strongly agree

5. I am motivated to learn.

Strongly disagree Disagree Undecided Agree Strongly agree

6. I explain what I need to do when I get an assignment.

Strongly disagree Disagree Undecided Agree Strongly agree

7. I list the strategies I'm using when I work on assignments.

Strongly disagree Disagree Undecided Agree Strongly agree

8. I check my progress towards achieving my goals.

Strongly disagree Disagree Undecided Agree Strongly agree

9. I modify (correct) my actions on my own to achieve my goals.

Strongly disagree Disagree Undecided Agree Strongly agree

(Appendices continue)

10. I modify (correct) strategies that are not helping me achieve my goals.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

11. I give helpful advice to my classmate on their work.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

12. I use comments from my teacher to improve on my work.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

13. I use comments from my classmate to improve on my work.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

14. I use comments from my family to improve on my work.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

15. I revise versions of my work to improve them.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

16. I reflect on the strategies I used to achieve my goals.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

17. I evaluate my own work (I look at my work to see if it is good or needs improvement).

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

18. I know how I am being evaluated.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

19. I make connections between the amount of time I spend on my work, and my achievement.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

20. I work well with other students.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
----------------------	----------	-----------	-------	-------------------

(Appendices continue)

SECTION 2: ePEARL USE

Please answer this section **ONLY** if you have used the ePearl software in your class

I liked using ePearl in my class because...

I did not like using ePearl in my class because...

ePearl helped me learn how to...

I would like to use ePearl again next year because...

I do not want to use ePearl again next year because...

What I liked the most about using ePearl is...

What I liked the least about ePearl is...

Thank you again for your collaboration!

Received December 15, 2011
Revision received July 25, 2012
Accepted November 12, 2012 ■

Universal Design for Learning and Elementary School Science: Exploring the Efficacy, Use, and Perceptions of a Web-Based Science Notebook

Gabrielle Rappolt-Schlichtmann
CAST, Inc., Wakefield, Massachusetts, and Harvard Graduate
School of Education

Samantha G. Daley, Seoin Lim, Scott Lapinski,
Kristin H. Robinson, and Mindy Johnson
CAST, Inc., Wakefield, Massachusetts

Science notebooks can play a critical role in activity-based science learning, but the tasks of recording, organizing, analyzing, and interpreting data create barriers that impede science learning for many students. This study (a) assessed in a randomized controlled trial the potential for a web-based science notebook designed using the Universal Design for Learning (UDL) framework to overcome the challenges inherent in traditional science notebooks, (b) explored how teacher characteristics and student use of supports in the digital environment were associated with productive inquiry science learning behaviors, and (c) investigated students' and teachers' perceptions of the key affordances and challenges of the technology to their learning. Use of the UDL science notebook resulted in improved science content learning outcomes ($\gamma = .34, p < .01$), as compared with traditional paper-and-pencil science notebooks, and positively impacted student performance to the same degree, regardless of reading and writing proficiency and motivation for science learning at pretest. Students of teachers with greater experience using science notebooks and students who more frequently used the contextual supports within the notebook demonstrated more positive outcomes. Students and teachers reported overall quite positive experiences with the notebook, emphasizing high levels of interest, feelings of competence, and autonomy.

Keywords: Universal Design for Learning, science notebook, elementary education, technology, design based research

Modern science education emphasizes learning that integrates higher order thinking skills with content-area knowledge in authentic problem-solving activities (Kame'enui & Carnine, 1998). Students are expected to learn actively through observation and interaction, rather than direct instruction (Chinn & Malhotra, 2002; Mastropieri, Scruggs, Boon, & Carter, 2001; National Research Council, 1996). Such active science learning requires students to develop and use a number of complex skills (Chinn & Malhotra, 2002), including making claims, observing, collecting information and data, analyzing, drawing conclusions, and presenting findings. Students need to build explanations by connecting their observa-

tions during inquiry science experiences to claims about what their observations might mean (McNeill & Krajcik, 2006; National Research Council, 2000; Sandoval & Reiser, 2004).

Within the mandate for active science learning, educators face a daunting set of challenges. The No Child Left Behind Act (2001) sharply increased accountability for raising the science achievement of all students (Individuals with Disabilities Education Act, 2004), and the student population is increasingly diverse. Although active science learning presents challenges for all students (De Jong & Van Joolingen, 1998; Keselman, 2003), the processes required may present particular difficulty for those who struggle with reading and writing (Englert, Raphael, Fear, & Anderson, 1988; Graham, 1990; Graham, Harris, MacArthur & Schwartz, 1991; Swanson, 1999; Wong, 2001), or who otherwise have low motivation for science learning (Keselman, 2003; Scruggs & Mastropieri, 1994).

Students may struggle not only with understanding the science concepts, and, in some cases, not even with the scientific inquiry process, but also with aspects of the active learning experience that are unintentional barriers to the deep learning being pursued. Sophisticated learning tools provide an opportunity to address these unintended, construct-irrelevant barriers. Such tools can assist teachers to more effectively and efficiently support students throughout the active science learning process and foster the skills and behaviors that are most productive in science learning.

In this article, we first describe a Universal Design for Learning (UDL) web-based science notebook designed to support elemen-

This article was published Online First September 9, 2013.

Gabrielle Rappolt-Schlichtmann, CAST, Inc., Wakefield, Massachusetts, and Harvard Graduate School of Education; Samantha G. Daley, Seoin Lim, Scott Lapinski, Kristin H. Robinson, and Mindy Johnson, CAST, Inc.

This research and development were supported by a grant from the Institute of Education Sciences. We thank the development team of the Universal Design for Learning science notebook, including Anne Meyer, Mindy Johnson, Kristin Robinson, Rick Birnbaum, Lisa Spitz, and Boris Goldowsky. We also thank our partners at Full Option Science Systems: Linda De Lucchi, Brian Campbell, and Larry Malone. Finally, and most importantly, we thank all of the children, families, and teachers who participated in this research.

Correspondence concerning this article should be addressed to Gabrielle Rappolt-Schlichtmann, CAST, Inc., 40 Harvard Mills Square, Suite 3, Wakefield, MA 01880-3233. E-mail: gschlichtmann@cast.org

tary school students and their teachers during active science learning. We then report the results of a study that (a) assesses the potential impact of this web-based science notebook to support improved content knowledge outcomes as compared with traditional paper-and-pencil science notebooks, (b) explores the factors that contribute to students' effective active science learning behaviors in the web-based notebook environment, and (c) investigates students' and teachers' perceptions of the key affordances and challenges of the technology. We provide this as a demonstration of the potential for UDL technology that provides options to address the variability in knowledge, skills, and preferences within an elementary school classroom. Such UDL technology can overcome barriers and support learning beyond the capabilities of more static learning tools.

Building ■ Better Science Notebook

Science notebooks are widely used to support the active science learning process and the development of scientific literacy (Hargrove & Nesbit, 2003; Klentschy, 2005), offering students the opportunity to engage in authentic scientific thinking (Hargrove & Nesbit, 2003) and providing teachers with opportunities for embedded formative assessment (Hargrove & Nesbit, 2003; Klentschy, Garrison, & Amaral, 1999; Shepardson & Britsch, 2004). When used effectively, science notebooks can support students to develop critical thinking and conceptual understanding (Keys, 2000; Miller & Calfee, 2004).

But, the research literature indicates that science notebooks are typically and primarily used in a mechanical way to record data, procedures, or definitions—and rarely to support development of deep understandings through the active science learning process (Baxter, Bass, & Glaser, 2001; Ruiz-Primo, Li, Ayala, & Shavelson, 2004). Furthermore, science notebooks present multiple barriers to students who struggle in the learning process because they require relative proficiency in reading and writing to be useful. Without high enough skill levels in these domains, students are unable to use notebooks to support the development of deep understandings through activity-based learning.

Designing a science notebook that incorporates the principles of UDL provides an opportunity to address both potential pitfalls in the use of science notebooks—UDL focuses on supporting the deep learning process and overcoming unnecessary barriers to such learning. UDL is a transdisciplinary framework that facilitates interaction between researchers from the learning sciences and professionals within education, focused on problems of common interest. The UDL framework can be used to reach a holistic understanding and work toward innovative solutions (Rappolt-Schlichtmann, Daley, & Rose, 2012; Rappolt-Schlichtmann & Watamura, 2010; Samuels, 2009).

The basic premise of UDL is that barriers to learning occur in the interaction with curriculum—they are not inherent solely in the capacities of the learner. Just as universally designed buildings provide options that accommodate a broad spectrum of users, universally designed tools and curricula offer a range of options for accessing and engaging with learning materials. “Universal” does not mean “one size fits all”; rather, it implies that curricula and materials are conceived of and designed to accommodate the widest possible range of learner needs and preferences.

Advances in technology have made the development of UDL approaches, texts, content curricula, strategy-based interventions, and assessment possible (Dolan, Hall, Banerjee, Chun, & Strangman, 2005; Rose & Meyer, 2002, 2006). Developed under the UDL framework, digital environments provide the necessary infrastructure and flexibility to allow for the creation of accessible, highly effective apprenticeship environments where students are actively guided in the process of constructing meaning through the provision of just-in-time feedback and contextual supports that can be gradually withdrawn as student expertise increases (Cognition and Technology Group at Vanderbilt, 1993; Collins, Brown, & Newman, 1989; Palinscar, 1986, 1998; Palinscar & Brown, 1984). Through this kind of design approach, teachers can be supported and provided with the flexible tools they need to create more effective and differentiated learning experiences for students (Dalton, Pisha, Eagleton, Coyne, & Deysher, 2002).

The Universally Designed for Learning Science Notebook (UDSN)

Like traditional science notebooks, the UDSN provides students with (a) space to collect, organize, and display observations and data; (b) space to reflect and make sense of inquiry experiences; and (c) multiple opportunities to demonstrate understanding and receive formative feedback. But with UDL as the design framework (CAST, 2011) and the potential of digital technology as the platform, the UDSN differs from traditional science notebooks in several key ways.

First, the UDSN was designed with a purposeful focus on lowering construct-irrelevant barriers to science learning. It was thus developed according to accessibility guidelines from the World Wide Web Consortium (W3C-WAI, 1999), Section 508 of the Rehabilitation Act (29 U.S.C. 794d), and the National Center for Accessible Media (2006). Text-to-speech technology is built directly into the notebook interface, as well as word-by-word English-to-Spanish translation, alt text and long descriptions for images, all actions are keyboard accessible, and a multimedia glossary is provided to provide just-in-time support for vocabulary use and development (see Figure 1). These features overcome such barriers for the many students whose literacy skills would interfere with the efficacy of materials that depend on proficiency in reading and writing (Hsu, 2004; Klecan-Aker & Caraway, 1997; Scruggs, Mastropieri, Bakken, & Brigham, 1993; Storch & Whitehurst, 2002; B. Y. White & Frederiksen, 1998), as well as for those students who use accessibility features due to sensory or motoric limitations, those for whom proficiency in English is a barrier, and others who would more effectively learn through use of built-in accessibility features.

Access to materials and tools, as is provided through the features described above, is an important advantage that can be built into digital technologies, but UDL offers another level of design advantage—access to learning. UDL places a premium on the use of contextual support (CAST, 2011) that, in this case, is intended to develop and then reinforce effective science learning behaviors. Pedagogy is built into the interface design itself, guiding students and teachers in the process of active science learning, and specifically the effective use of science notebooks. For example, the navigation of the UDSN (see Figure 1A) provides a conceptual anchor in both words and

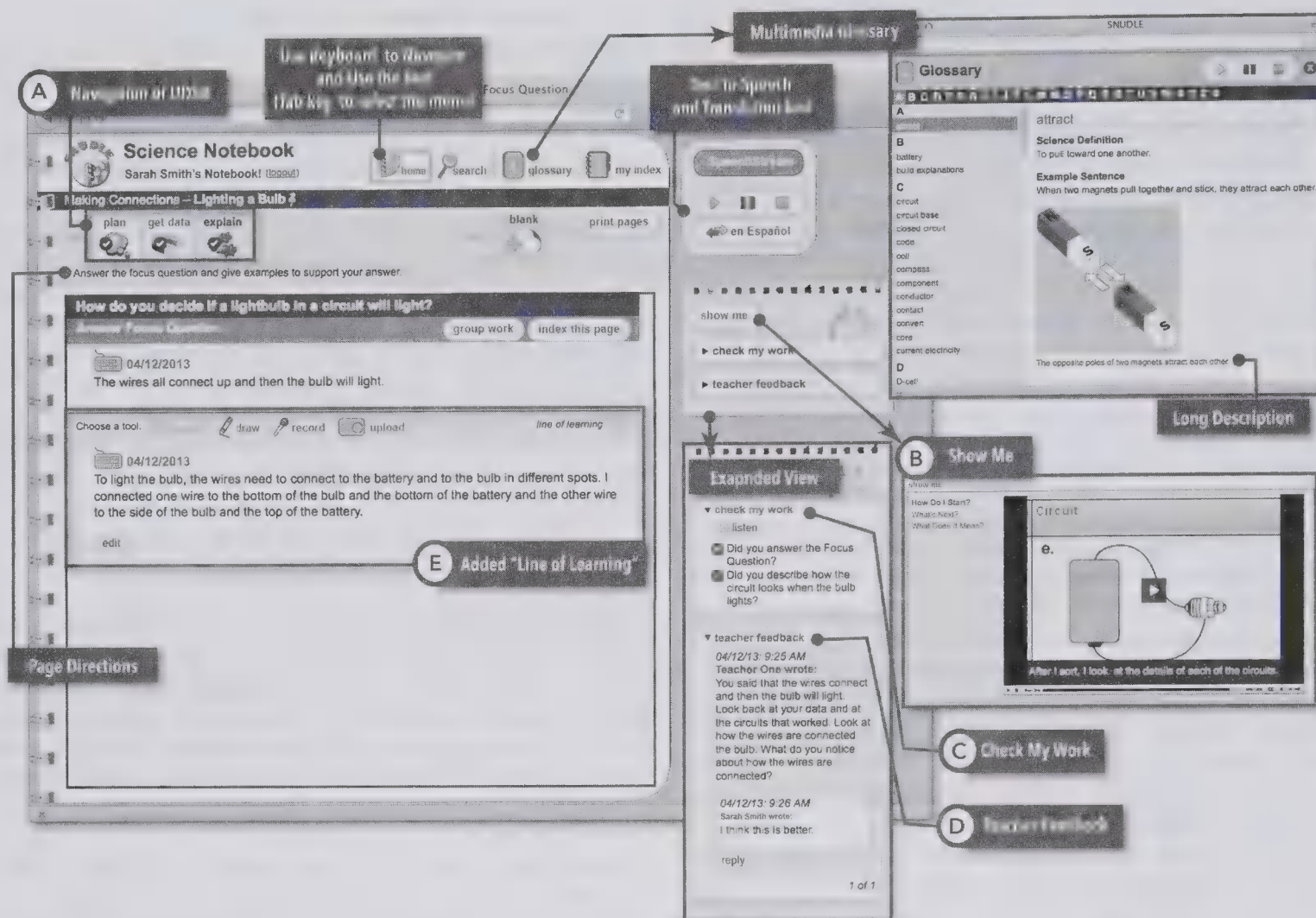


Figure 1. Student interface, Universally Designed for Learning Science Notebook.

pictures. As students complete each part of a science activity, they are reminded through the navigational structure of the UDSN that they are moving through a process: plan, get data, explain. Once students begin to build an explanation for their inquiry experience, they are further provided with contextual supports to facilitate, guide, and then reinforce the process behaviors necessary for effective science notebook use. The "Show Me" feature (see Figure 1B) provides brief captioned videos, where students are guided in how to go about building an explanation—students are prompted to think about "How do I start?" "What's next?" and "What does it mean?" Students can use the "Check My Work" support (see Figure 1C) to be sure their explanations contain all of the necessary components. Among other things, they are prompted to think about making direct reference to their data and observations and to use relevant vocabulary from their inquiry experiences. Furthermore, students can choose to express their thinking through any of a variety of multimedia response options, including typing, drawing, audio recording, or uploading a picture (see Figure 2).

In addition to incorporating pedagogy in the student-facing interface, active science learning is facilitated through the role of the teacher in using the UDSN. Teachers are supported in the process of providing feedback to students (see Figure 3). Feedback is a necessary catalyst for self-regulated, effortful, and

persistent learning behavior, and formative assessment is needed so that teachers can be successful at differentiating the affordances of the UDSN tool to student strengths and weaknesses (Butler & Winne, 1995). In the UDSN teachers can view all of their students' explanations at once or one at a time, and make quick notes to themselves about whether the student "got" the concept or not (see Figure 3A). Teachers are provided "What to look for" information (see Figure 3B), including core concepts, common misconceptions, and model feedback. "Teacher timesaver" (see Figure 3C) provides sentence starters for feedback that is process oriented and catalogues recent feedback the teacher has given to other students. Teachers are prompted and supported to provide feedback that may include corrective information, alternative strategies, information to clarify ideas, or encouragement to engage in the scientific process (Hattie & Timperley, 2007).

Students can then use teacher-provided feedback to pursue active science learning and engage in productive science learning behaviors. In response to feedback, students are prompted to revisit observations (see, e.g., Figure 1D) to revise their explanations and then add a "Line of Learning" to their notebook (see Figure 1E). When students add a line of learning, they are indicating a shift in their thinking in response to feedback or new information gathered through inquiry experiences. In this way they

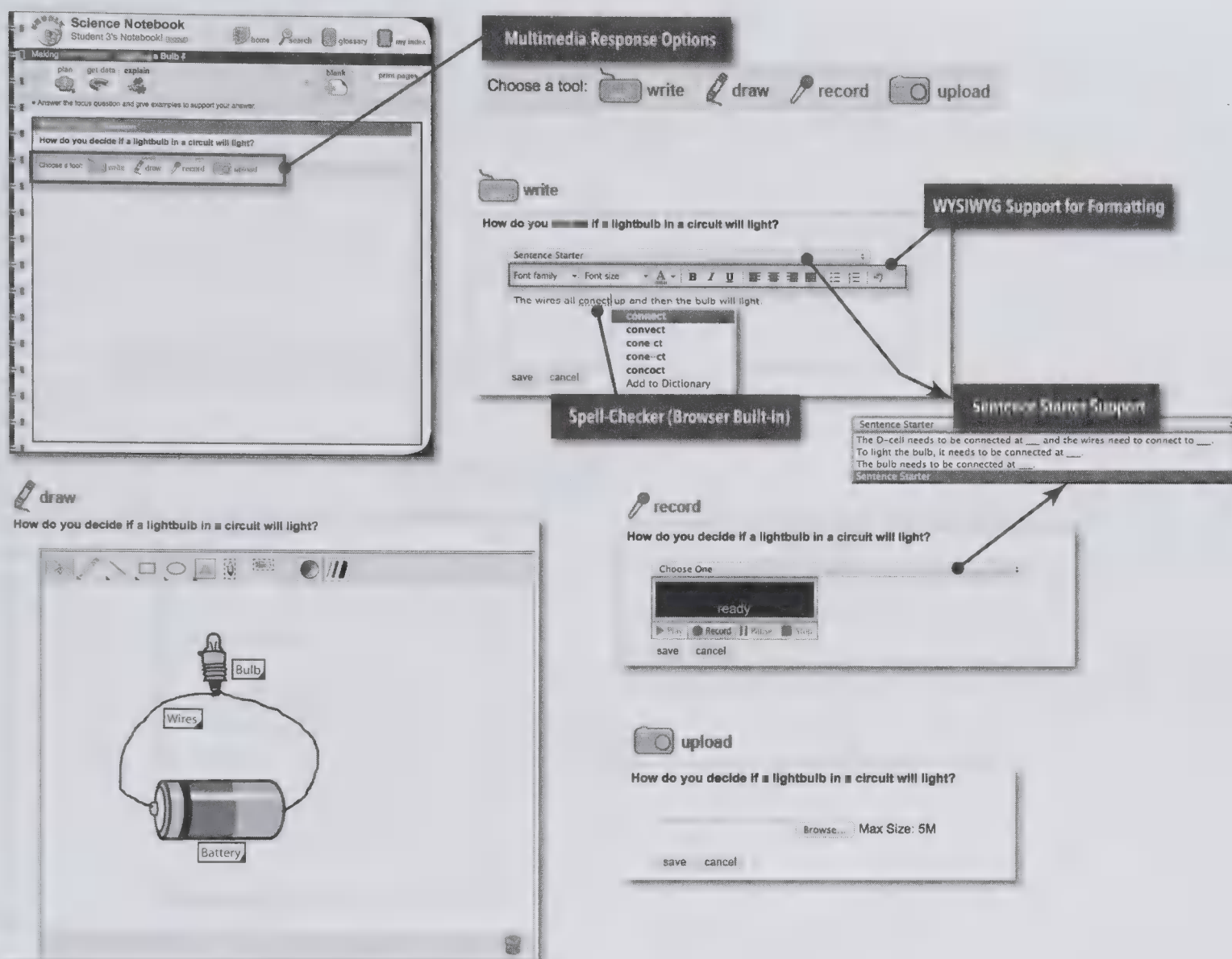


Figure 2. Multimedia response options and supports for students, Universally Designed for Learning Science Notebook.

revise their ideas but retain the string of explanations they have built—the evolution of their thinking is evident and emphasized in the interface of the UDSN.

Method

The UDSN was developed through a process of progressive refinement using design-based research methodology. Design-based research is a formative evaluation approach to intervention and technology development where the goal is to refine both existing theory from the research literature and to generate “usable knowledge” to improve educational practice (Collins, Joseph, & Bielaczyc, 2004; Flagg, 1990; Reeves & Hedberg, 2003). Within this framework, development and research take place through continuous cycles of design, implementation, analysis, and redesign (Cobb, 2001; Collins, 1992). The current study represents the summative research associated with this work (Rappolt-Schlichtmann, Daley, Lim, Robinson, & Johnson, 2011).

The overarching goal of the research was to determine whether the UDSN enhanced the science learning of diverse fourth-grade

students in authentic public school settings relative to paper-and-pencil science notebooks and to understand student and teacher experiences in using the web-based science notebook at the fourth-grade level. Furthermore, we sought to quantitatively and qualitatively explore the mechanisms by which the UDSN operates on student science learning. The following research questions were posed:

- Research Question (RQ) 1 (Overall Impact): On average, do students in classrooms using support-rich, UDL science notebooks learn and understand more about science than similar students in similar classrooms using traditional paper-and-pencil science notebooks?
- RQ 2 (Differential Impact): On average, is the impact of the UDSN the same for students at various reading and motivation levels?
- RQ 3 (Use): Do students use the UDSN in ways that would indicate productive science notebook use?
- RQ 4 (Differential Use): Do students whose teachers have more professional experience and students who more frequently

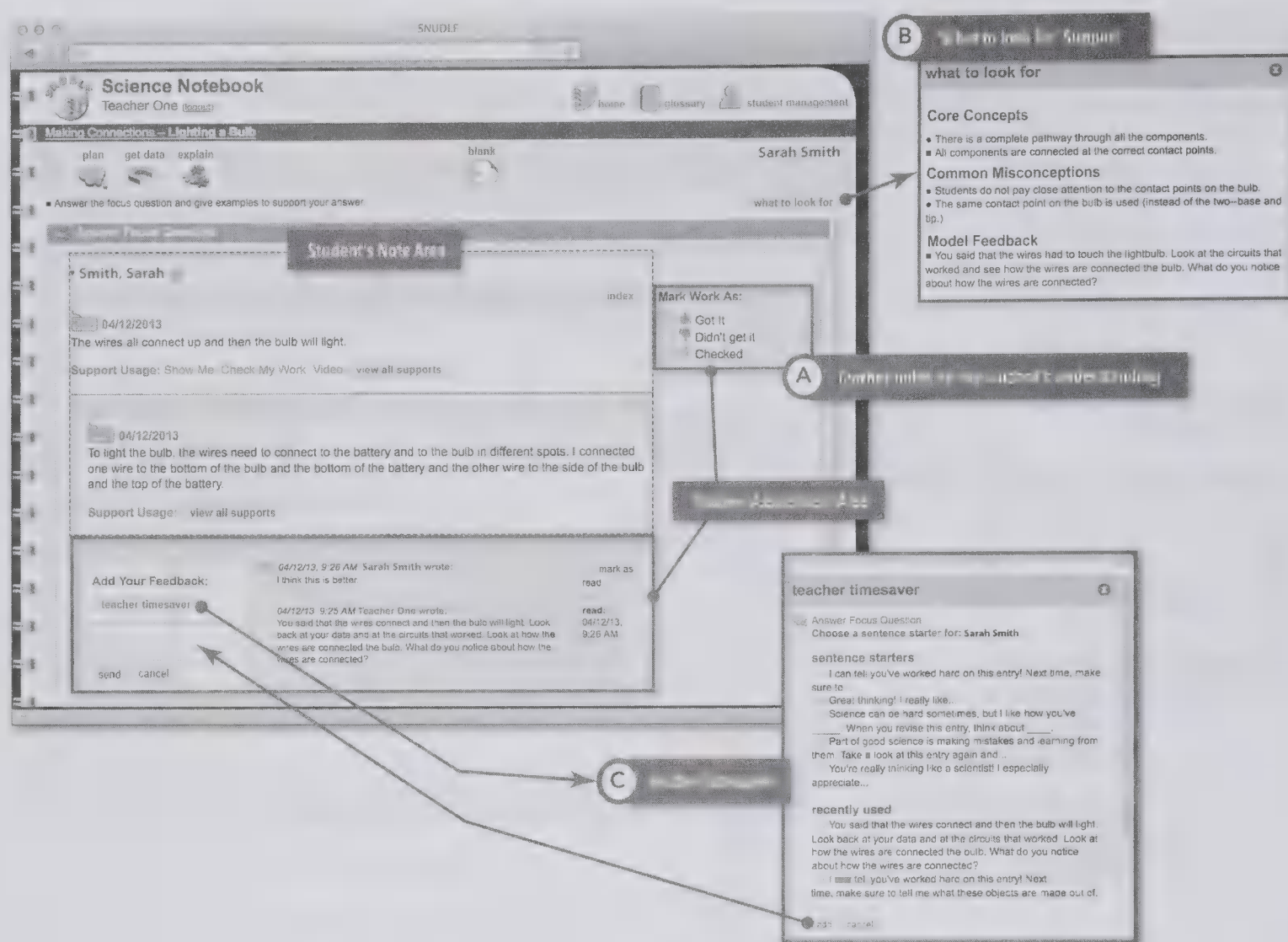


Figure 3. Teacher interface, Universally Designed for Learning Science Notebook.

use the contextual process-focused supports in the UDSN tend to engage in more productive science notebook learning behaviors?

- RQ 5 (Perceptions): What are students' and teachers' perceptions of the usefulness of the UDSN in science learning?

Site and Participants

The data were collected from eight schools within a large southeastern school district in the United States. The district consisted of rural, suburban, and urban schools. The sample consisted of 621 fourth-grade students from 28 different classrooms and 22 teachers. Overall, the student sample consisted of nearly 35% minority students, and 10% of the students were characterized as having an Individualized Education Program (IEP), or a 504 plan.

Procedure

To answer RQ 1–4, we conducted a randomized controlled trial that investigated the potential of the notebooks to support the development of student content knowledge, as compared with traditional science notebooks. Over the course of the 8–10 weeks of the study, fourth-grade students used either the UDSN notebooks (treatment group) or traditional science notebooks (control

group) as they completed the magnetism and electricity unit of the Full Option Science System (FOSS) curriculum. One of the most widely used in the United States, FOSS is a research-based, K–8 science curriculum developed by the Lawrence Hall of Science (Banilower, 2002). Students and teachers regularly use paper-and-pencil science notebooks as a part of this curriculum.

Participating teachers were randomized to either the treatment or the control group in two steps. First, pairs of teachers within each school were identified using a matching index of teacher experience and classroom demographics. Second, from the pairs, we randomly assigned one teacher to the control group and one to the treatment group. All teachers who participated in the research study, regardless of group assignment, received a 1-day FOSS training session focused on the use of science notebooks to support science learning at the elementary school level and a 1-hr high-level orientation to the UDL framework. Teachers in the treatment group also participated in an additional 2 hr of training, where they were made aware of the additional features the UDSN provided above and beyond paper-and-pencil notebooks, and had an opportunity to practice using those features as if they were students. Teachers in all conditions received follow-up support 1 and 2

weeks after beginning the implementation, and then technical support for the remainder of the implementation.

Prior to beginning the intervention, students in the treatment condition were given training on the use of UDSN by one of two project team members. This included an overview of the UDSN features, as well as a .5-hr session during which students were able to try the UDSN and ask questions. The project team also spent 1 week within each school ensuring the technology (computers and Internet access) was sufficient for implementation of UDSN. In one instance when technology was a barrier, we provided the classroom with an extra router to allow the use of a laptop cart.

To answer RQ5, we conducted focus groups with students and interviews with teachers who used the UDSN during the summative research study described above. Eighty-four students with experience using the UDSN (six students from each of the 14 experimental classrooms) were selected to participate in focus groups. The selection process was purposeful so as to include both high- and low-achieving students, students with and without IEPs, and an equal number of boys and girls. All 11 teachers who implemented the UDSN as a part of the summative study were interviewed. Focus groups with students and interviews with teachers were conducted within 2 weeks of the conclusion of the study period and lasted 60–75 min. Students and teachers were asked the same set of semistructured questions (e.g., for students, “Do you think your UDSN helped you learn science or not?; How?”; “Were there times when your UDSN made it more difficult or got in the way of your science learning?”; “If you think the UDSN helped you learn science, describe something specific you did in your UDSN that helped you learn science”; “What about your UDSN was frustrating or difficult?”), but the format of these sessions was open-ended so as to allow for the spontaneous elaboration of thoughts. Focus groups and interviews were taped and then transcribed for later coding.

Measures

Magnetism and electricity content knowledge (Assessing Science Knowledge [ASK]). The ASK Survey (Ferguson, Long, & Kennedy, 2009) was used to measure changes in knowledge of science content taught within the FOSS curriculum. Student proficiencies were computed using a maximum likelihood estimation method using ConQuest software. Student proficiencies are computed as the location on a content-specific scale that is most probable for students given their responses to the items. Because student proficiency estimates are not provided to teachers directly, we did not scale the raw logit scores onto an external positive scale such as 0–100. Thus, student proficiency estimates ranged between -6.0 and 6.0 .

Motivation for science (Motivation for Science [MFS] Inventory). Due to the lack of published assessments on student motivation in science, we developed the MFS inventory. After an extensive literature review, we derived questions representing four constructs key to student motivation at the elementary level: self-efficacy, interest, desire for challenge, and social behavior. The survey consisted of 23 Likert-type items (“I am good at science,” “I like science when it is hard and challenging”). Students rated items on a scale ranging from 1 (*very different from me*) to 4 (*a lot like me*). Questions related to the “social” construct did not reliably hang together and so were eliminated during the pilot test phase of

this project. Reliability for the pilot test sample of the MFS survey was .85; for the experimental sample, it was .89.

Reading and writing proficiency (Measure of Academic Progress [MAP]). The district administered the MAP to assess students’ proficiency in reading and writing; scores were collected with permission from participants. The MAP assessments are computerized-adaptive tests developed by the Northwest Evaluation Association (NWEA) and used in 2,570 school systems across 49 states nationwide to assess student proficiency in a variety of areas including English Language Arts. The MAP tests were developed from large pools of items that have been calibrated for their difficulty. Prior research indicates the MAP assessments are valid and reliable indicators of student proficiency in targeted areas (NWEA, 2005).

Electronic usage log. The web-based UDSN includes an electronic usage log. This log allows researchers to view items clicked, including use of navigation, supports, choices of whether to type, draw, or audio-record responses, and other patterns of use in time series within the program for both students and teachers. Using this log, we can determine the frequency of use and usage patterns, helping us to determine in what way components are being used, thus informing design, organization, and content.

Teacher background. Teacher background characteristics were collected via questionnaire. This information included age, years of teaching experience, years of teaching at the current grade level, years teaching with the FOSS curriculum, years teaching using science notebooks, number of hours of science-focused professional development in the past 5 years, and hours spent teaching science each week.

Results

Quantitative Analysis

To answer RQ 1–4, we used a multilevel modeling approach. Multilevel modeling can account for measurement and sampling error when variables operate at different but related levels of organization, resulting in correctly adjusted standard errors for the treatment effect (Raudenbush & Bryk, 2002; Singer, 1998). To answer the first two research questions, we fit a series of three-level models in which students were clustered within teachers, and teachers were clustered within schools. To answer RQ 3 and 4, we fit two-level models (students within schools) because these questions deal only with the treatment group, making the classroom and school levels of analysis synonymous due to our sampling and randomization procedure. Continuous covariates were grand-mean centered, whereas categorical variables were represented as 0/1 indicators. Goodness of fit was assessed using the -2 log-likelihood and Akaike’s information criterion statistic where smaller is better. Variables were systematically added to the model and then maintained if significant. All models were fit using either the PROC MIXED procedure within SAS or xtmixed within Stata.

RQ 1 (Overall Impact). On average, do students in classrooms using support-rich UDL science notebooks learn and understand more about science than similar students in similar classrooms using traditional paper-and-pencil science notebooks? We list the sample means and standard deviations for all variables included in our analyses in Table 1. Note that at posttest, content knowledge was relatively higher among students in the treatment

Table 1
Mean Performance Scores and Covariates of Students in the Treatment and Control Groups

Variable	Treatment			Control		
	<i>n</i>	<i>M (SD)</i>	Range	<i>n</i>	<i>M (SD)</i>	Range
Outcome (posttest)						
M&E Knowledge	355	.42 (.9)	-1.8-3.7	168	.01 (.9)	-4.6-2.8
Predictors (pretest)						
M&E Knowledge	355	-1.4 (.7)	-4.6-.61	168	-1.7 (.9)	-4.6-.61
MAP, Language Arts	355	217 (10.8)	172-280	168	215 (12.4)	163-243
Motivation for Science	346	41.5 (7.7)	16-52	168	40.6 (7.3)	19-52

Note. M&E knowledge = Magnetism and Electricity Knowledge; MAP = Measure of Academic Progress.

group ($M = .42$, $SD = .9$) than in the control group ($M = .01$, $SD = .9$). Though the standard deviation from the mean for content knowledge at posttest was the same for both groups, the range of values in the treatment group (-1.8, 3.7) was smaller as compared with the control group (-4.6, 2.8). Mean values and standard deviations between treatment and control on the MAP language arts test, motivation for science survey, and magnetism and electricity content assessment were similar at pretest, though values on the content knowledge pretest were slightly higher in the treatment group ($M = -1.4$, $SD = .7$) as compared with control ($M = -1.7$, $SD = .9$). We controlled for pretest content knowledge in our model of impact.

In Table 2, we present the baseline control and final model, as well as a series of models describing interaction tests between covariates and treatment extracted from the larger taxonomy of models that we fit systematically during data analysis. We present estimates and goodness-of-fit statistics for both the fixed effects and the variance components, along with goodness-of-fit statistics for the overall model. RQ 1 concerns the overall impact of the

UDSN. We find that students using the UDSN demonstrated greater knowledge of the science content at posttest than peers using traditional science notebooks ($\gamma = .32$, $p < .05$), controlling for pretest levels of content knowledge ($\gamma = .25$, $p < .001$), reading skills ($\gamma = .04$, $p < .001$), and motivation for science ($\gamma = .01$, $p < .05$) (see Table 2, Final column; overall model fit at $\chi^2[3] = 388$, as compared with baseline).

Interpreting the parameter for the fixed effect associated with treatment indicates that, on average, students in treatment classrooms exhibit proficiency scores .32 points higher than students in control classrooms. Forty-four percent of the explainable variation in the ASK posttest is explained by assignment to the treatment group in this model.

RQ 2 (Differential Impact). On average, is the impact of the UDSN the same for students at various reading and motivation levels? A pair of interactions was added individually to the overall impact model to determine whether differential effects of the UDSN were evident for students with varying levels of pretest reading skills and motivation for science. Neither the interaction

Table 2
Fixed Effects Estimates (Top) and Variance-Covariance Estimates (Bottom) for Models Describing Predictors of Knowledge of Magnetism and Electricity at Posttest, Including the Impact of Condition

Parameter	Baseline	Final	Covariates	Interaction 1	Interaction 2
Fixed effects-Level 1					
Intercept	.91 (.14)***	.43 (.16)*	.63 (.12)**	.43 (.16)*	.49 (.16)*
ASK pretest	.49 (.05)***	.25 (.05)***	.26 (.05)***	.25 (.05)***	.21 (.05)***
MAP		.04 (.003)***	.04 (.003)***	.04 (.006)***	.04 (.003)***
Motivation		.01 (.004)*	.01 (.004)*	.01 (.004)*	.01 (.008)
Fixed effects-Level 2					
Treatment		.32 (.13)*		.32 (.14)*	.32 (.13)*
Treatment \times MAP				.004 (.01)	
Treatment \times Motivation					-.002 (.01)
Random effects					
Between schools	.03 (.06)	.05 (.06)	.01 (.06)	.05 (.06)	.05 (.06)
Between teachers	.16 (.07)*	.05 (.03) [†]	.09 (.06) [†]	.05 (.04) [†]	.04 (.03) [†]
Within teachers	.55 (.03)***	.39 (.03)***	.39 (.03)***	.39 (.03)***	.39 (.03)***
Goodness of fit					
-2LL	1,219	831	834	839	833
AIC	1,225	837	840	845	839

Note. Multilevel modeling (students, within teachers, within schools) was used to estimate the effects of treatment (Raudenbush & Bryk, 2002). All models were fit using the PROC MIXED procedure within SAS. Standard errors are in parentheses. ASK pretest = Full Option Science Systems Assessing Science Knowledge Pretest for Magnetism and Electricity; MAP = Measures of Academic Progress for Language Arts; Treatment = teacher random assignment to Universally Designed for Learning Science Notebook (1) or Traditional Science Notebook (0); Motivation = Motivation for Science Survey; -2LL = -2 log-likelihood; AIC = Akaike's information criterion.

[†] $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

between treatment and pretest reading skills on the MAP assessment ($\gamma = .004, p > .05$) nor the interaction between treatment and motivation for science ($\gamma = -.002, p > .05$) was statistically significant or improved the overall fit of the model (see Table 2, Interaction 1 and 2, respectively). Thus, the relationship between condition and posttest score did not differ by reading level, nor did the relationship between condition and posttest score differ by students' levels of motivation for science.

RQ 3 (Use). Do students use the UDSN in ways that would indicate productive science notebook use? To answer RQ 3, we conducted an exploratory analysis of student use of the UDSN (treatment condition only). We constructed metrics that describe student behaviors derived from click-by-click usage signals collected via the electronic usage log. Given the UDL framework and the goals of our design process, and after looking to the research literature, we were able to identify four behaviors that could be extrapolated from student clicks catalogued in time series: (a) number of sessions using the UDSN (*overall use*), (b) number of completed entries in which students were asked to explain key science concepts (*reflective consolidation and demonstration of knowledge*), (c) reviewing data/observations when asked to construct explanations of inquiry experiences (*use of data or inferences from observations*), and (d) revision of previous entries to reflect new understanding or teacher feedback (*continuous learning and recursion on ideas*).

In Table 3, we provide descriptive statistics for these four behaviors, totaled across the entire implementation period, averaged by number of sessions each student used the UDSN, and, where appropriate, averaged across the number of key concept explanations completed. This set of descriptive statistics provides both an overall picture of use across the 8-to 10-week implementation period and an idea of how often each discrete behavior was evident for each student.

As Table 3 indicates, students used the UDSN an average of 10 times across the implementation period, or approximately once per week. This varied, however, with some using it only a handful of times, and others using it 2 or 3 times the average amount. Students answered the key explanation questions an average of 12 times, or, again, just over once per week. But, the range in number of responses approximated the range in total number of UDSN sessions across the implementation period, and the number of key concept entries averaged to once per session.

Revisions to previous notebook entries and reviewing data in order to answer the key concept questions were somewhat more

discrete process behaviors that reflected not only what was completed but also how students went about completing their work in the UDSN. On average, students revised about four notebook entries over the course of the entire implementation period, but the variation in this activity was substantial. Although almost 20% of students never revised a post, almost 10% of students revised more than 10 posts. Students overall averaged a revision to a post in about one out of three UDSN sessions, and this was almost equivalent to the average per key concept question; most revisions to previous posts were revisions to responses to the key concept questions. Students averaged 36 instances of looking at a data-focused page in the UDSN before completing an answer to a key concept question, with a dramatically wide range from no instances to 149. Ten percent of students logged 69 or more instances of this key science process behavior, and the average was three per session and per key concept entry.

RQ 4 (Differential Use). Do students whose teachers have more professional experience and students who more frequently use the contextual process-focused supports in the UDSN tend to engage in more productive science notebook learning behaviors? The variability in students' science notebook learning behavior described above makes clear the need to understand what characteristics potentially promote such productive behaviors. We considered two sets of potential characteristics: the experience level of the student's teacher and the frequency with which the student engaged with the relevant supports in the UDSN web-based environment designed to promote the emergence of such behaviors. The four productive behaviors described above (number of UDSN sessions, number of key concept questions answered, number of edited posts, and number of reviews of data when answering key concept questions) were highly correlated with each other (r ranged from .56 to .74; all $ps < .001$); the total of these behaviors across the implementation period was combined in a single composite using a principal components analysis for the purpose of data reduction. All four behaviors loaded on a single factor.

Using multilevel models in which students are clustered within teachers, we fit a set of models examining each type of characteristic predicting the composite outcome of productive behaviors. We present the results of the taxonomy of models in Table 4. The three teacher experience variables of interest were years of experience using notebooks in science instruction, years of overall teaching experience, and years of experience with the FOSS curriculum. The goodness of fit of each model was compared with a baseline control model, including pretest content knowledge.

Table 3
Mean Frequency of Students' Productive Notebooking Behaviors Across the Intervention Period, Averaged Across Number of UDSN Sessions, and Averaged Across Number of Explanations of Key Concepts Entered

Variable	n	Total across intervention period		Average per UDSN session		Average per key concept explanation entered	
		M (SD)	Range	M (SD)	Range	M (SD)	Range
UDSN sessions	411	10.48 (4.3)	1–30	—	—	—	—
Key concept explanations entered	411	11.52 (6.2)	0–33	1.07 (.43)	0–2	—	—
Revisions to previous notebook entries	411	3.86 (4.3)	0–33	.32 (.26)	0–2	.32 (.29)	0–3
Reviewing data when entering key concept explanations	411	35.78 (24.8)	0–149	3.38 (2.2)	0–19	3.20 (1.9)	0–15

Note. UDSN = Universally Designed for Learning Science Notebook. Dashes indicate that data are not applicable.

Table 4

Fixed Effects Estimates (Top) and Variance–Covariance Estimates (Bottom) for Models Describing Predictors of Process Behaviors Using the UDSN

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed effects-Level 1						
Intercept	32.91 (3.67)***	25.47 (4.33)***	25.02 (7.72)**	28.87 (5.87)***	29.58 (3.85)***	22.50 (4.49)***
ASK pretest	3.06 (1.04)**	3.03 (1.04)**	3.11 (1.04)**	3.08 (1.04)**	2.67 (1.05)*	2.654 (1.04)*
Use of process supports					0.27 (3.85)**	0.262 (.10)**
Fixed effects-Level 2						
Years of notebook use		1.307 (.53)*				1.27 (.53)*
Years of teaching experience			0.57 (.49)			
Years using FOSS curriculum				0.54 (.61)		
Random effects						
Between teachers	10.13 (2.55)***	7.94 (2.12)***	9.84 (2.70)***	10.22 (2.76)***	10.02 (2.54)***	7.99 (2.22)***
Within teachers	13.05 (.50)***	13.05 (.50)***	13.05 (.50)***	13.05 (.50)***	12.94 (.50)***	12.94 (.50)***
Goodness of fit						
–2LL	2,828	2,823	2,827	2,827	2,825	2,819
AIC	2,837	2,833	2,837	2,837	2,835	2,831

Note. UDSN = Universally Designed for Learning Science Notebook; ASK pretest = Full Option Science Systems (FOSS) Assessing Science Knowledge Pretest for Magnetism and Electricity; –2LL = –2 log-likelihood; AIC = Akaike's information criterion.

* $p < .05$. ** $p < .01$. *** $p < .001$.

The parameters for Model 2 indicate a statistically significant, positive effect of teachers having had more experience using science notebooks in instruction ($\gamma = 1.307$, $p > .05$); controlling for pretest content knowledge, students of teachers with one more year using science notebooks, on average, exhibit 1.3 more productive behaviors. Neither years of overall teaching experience (Model 3) nor years of experience with the FOSS curriculum (Model 4) were statistically significant predictors of student productive behaviors.

As was the case for the productive behaviors used as the outcome in these analyses, the use of supports intended to encourage these behaviors in the UDSN were also highly correlated with each other. Students who used the “check my work” coach also tended to access agents to get guidance for understanding questions or answering questions and tended to open videos designed to provide guidance on scientific thought processes. R values between the three ranged from .33 to .53, all with $p < .001$. Like the outcome behaviors, we combined total use of these supports into a single composite using principal components analysis for the purpose of data reduction in which all loaded on a single factor.

As shown in Model 5 of Table 4, the composite of use of process-focused supports in the UDSN was a positive and statistically significant predictor ($\gamma = .27$, $p > .01$) of productive science notebook behaviors; controlling for pretest content knowledge, students who used the contextual supports more frequently were more likely to engage in the desired outcome behaviors. In Model 6, we provide a final model combining the two substantive predictors of the teacher's years of experience using science notebooks in instruction ($\gamma = 1.27$, $p > .05$) and students' use of contextual, process-focused supports in the UDSN ($\gamma = .26$, $p > .01$); both remain positive and statistically significant, controlling for student pretest content knowledge ($\gamma = 2.65$, $p > .05$).

Qualitative Analysis

To answer RQ 5 (what are students' and teachers' perceptions of the usefulness of the UDSN to science learning?) a grounded

approach to qualitative data analysis was used. Because the UDSN is the first web-based program of this kind, we felt it was important to stay as close to the data as possible, allowing for the emergence of unexpected themes and serendipitous findings. Student focus groups and teacher interviews were coded and analyzed by the first author and two research assistants using the constant comparative method first developed by Glaser in 1965 (Glaser & Strauss, 1967). An inductive approach was adopted whereby categories and thematic connections were identified within germane units of data through a reductionist process (Creswell, 2007; LeCompte & Preissle, 1992; Miles & Hubberman, 1984).

An electronic database was generated using the qualitative analysis software NUD*IST nVivo (2002). Conceptually salient comments were marked with a series of *codes* and then extracted from the text. Codes were grouped into *concepts*. *Thematic categories* were formed. To validate the formation of concepts and thematic categories, transcripts were scoured for negative cases and disconfirming evidence. Once all relevant data were grouped into thematic categories (saturation), hypotheses describing student and teacher perceptions as to the usefulness of the UDSN in science learning were formed. Five thematic categories emerged from the data analysis. Four of these categories dealt expressly with engagement as students were or were not motivated to engage in the processing of their inquiry science experiences, whereas the fifth dealt with practical challenges associated with the use of the UDSN in comparison to paper-and-pencil science notebooks.

Interest, excitement. Without exception, student focus groups and teacher interviews noted high excitement and/or interest among students in using the UDSN, with students also reporting that the UDSN was more “fun” than paper-and-pencil science notebooks. When students offered an explanation as to why the UDSN was more “fun,” they most often noted that the UDSN reflected a personal interest in technology. A few students noted that technology, although ubiquitous in their personal lives, was not abundant at school.

... because the [UDSN] is not just paper pages. Lots of kids love electronics, like I do. At home it's like cell phones and computers and TV. It's more fun for me to get on the computer and work instead of writing with pencil. [Student Comment]

I think they [students] were way more engaged. They were more energized. It was exciting to them. And I really was kind of thinking that it might wear off a little bit as they got into it, but I don't think it did at all. [Teacher Comment]

As illustrated in the quote above, teachers typically noted that the excitement level did not wane over the 8- to 10-week implementation period despite significant challenges related to hardware and broadband availability.

Doing science, going deeper. Within the structured component of the interviews and focus groups, teachers and students were prompted to think about and then explain how the UDSN did or did not help science learning. Again, most students began with the idea that the UDSN was more "fun" than paper-and-pencil science notebooks but then expanded to articulate that because the UDSN was more "fun," they spent more time "doing" science.

Almost everybody likes things being exciting. The [UDSN] was exciting and fun ... when it's fun, you want to do it longer and with [UDSN], you just really don't want to get off. You want to keep doing it. [Student Comment]

A student in one focus group reported that getting time on the computer to work on science using the UDSN became competitive (see below). Most focus groups reported that they found ways to use the UDSN outside of class time even though their teacher did not require it.

I think we [students] did a lot more text and detailed pictures because we want to stay on the [UDSN] longer. And if class was over, we save what we have and go and do the [UDSN] by ourselves, which was really fun. And if people said, hey, it's my turn, you've been on for like a whole class, it's like sorry, I'm doing my stuff. [Student Comment]

Taking ownership, showing science thinking. When asked what they spent more time doing using the UDSN, students most often reported that they worked on building their explanations or revising work as prompted by their teacher. Students noted that in comparison to paper science notebooks, they were more likely to attempt an explanation when using the UDSN: "Before we knew about [UDSN], we just did experiments ... we normally didn't explain much" [Student Comment].

These comments were confirmed in teacher interviews where eight of the 11 teachers interviewed noted that with the UDSN, they were able to see student thinking about science more clearly, and more often than when they were using paper notebooks. One teacher commented that this was the first time she had been able to see her students' original thinking about their science experiences in her class: "When they were working on the [UDSN], it was pretty much their [students] own original work and, so I was able to see not my thinking, but their thinking for the first time" [Teacher Comment]. Students expressed similar thoughts in four of the 14 focus groups. "It's not just the teacher telling you the answers, it's actually you getting into your work and actually doing it" [Student Comment]. Most teachers noted that their students seemed to have a greater sense of ownership over their work as represented within the UDSN.

It gave them more of the sense of ownership of their own [science note-] book so they were more willing to keep up with it and keep working it. Many of them asked me, 'I didn't get finished with that lab, can I log onto UDSN and do that?' And of course when they are doing their paper notebooks, you know, usually I have to say to students, 'You have to finish this!' and it's not them asking me 'Can I go in there and do this?'" [Teacher Comment]

As illustrated in this quote, most teachers interviewed had hypotheses as to why student effort and excitement remained high over the course of the implementation, as well as why student thinking about science was more accessible with the UDSN. In the quote reported above, the teacher articulates her idea that student's sense of ownership reciprocally contributed to and reinforced high levels of interest, excitement, and effort.

Feeling competent, showing what you know. Without exception, student focus groups indicated that they liked having access to the UDSN contextual supports (e.g., Figure 1B and Figure 1C; Figure 2 sentence starter supports) and that the presence and use of these supports made them feel more confident and competent in their work. For example, several students commented that sometimes they have trouble understanding what their teacher is asking them to do and expressed satisfaction with having access to guidance on inquiry activities through the UDSN. "My teacher sometimes tells us to do something, and I don't really understand, so on the [UDSN] they have videos that tell us how to do it [inquiry activity] and that was great" [Student Comment]. Whereas another student noted, "I liked [UDSN] because if you ever got stuck on something and you forgot what to do, you could just click on the help thing on the side and it would like appear and help you" [Student Comment]. One student described the UDSN as a resource to be leveraged in the process of doing his science work: "The best thing about the [UDSN] is that it's like a backup resource" [Student Comment].

Although all of the student focus groups commented on the utility of various contextual support features of the UDSN, the reported usefulness of specific supports across students was highly varied. Some students placed high value on mechanical supports like spell check (see Figure 2), whereas others focused on conceptual or organizational supports. All of the focus groups commented positively on the sentence starter supports (see Figure 2) and offered to help students begin their explanations.

I thought if you really didn't know how to start your explanation, I thought it kind of helped you get an idea of how to make the words. The sentence starter, it gives me ideas and I just ... I just say, 'Hey! This could be easy!' whenever I think it's hard." [Student Comment]

Likewise, all of the teachers interviewed reflected on the utility of the contextual supports in facilitating student independence in their inquiry science work and science learning. One teacher hypothesized that the presence of the contextual supports was anxiety reducing for her students, "I think having all the resources in the [UDSN] took some of the anxiety out of it for them [students]. Just knowing that they [students] had help there even if they didn't use it, they knew they could" [Teacher Comment]. Half the teachers interviewed reported that the UDSN contextual supports and organizing structure were helpful to their practice,

I think it also made me more accountable in some ways. Just because I used it on my SMART Board with the [UDSN] so much that it kept

me more structured. It kept me focused, I was more in control of the lesson. [Teacher Comment]

All of the teachers interviewed and each of the student focus groups noted the challenges of the paper-and-pencil format for students in communicating what they know at the fourth-grade level,

[I]n a paper-and-pencil notebook, you know, there are a lot of kids who are challenged with handwriting and, honestly, spelling is atrocious, so when they go back to re-read something or use their notebook to study, they don't even know what they wrote. [Teacher Comment]

Nine of the 11 teachers interviewed indicated that the response options (see Figure 2) were essential to the utility and effectiveness of the UDSN.

The best thing about the UDSN is that there are so many ways kids can do things—you can upload pictures, you can draw, can record, you can type and translate. That's the multiple options of expressing their learning. It helped me realize that some students who were lower were getting it more than I thought, and then I could focus my teaching more productively. [Teacher Comment]

One teacher noted that the multiple response option feature of the UDSN was especially important for the students in her class with IEPs:

for me to hear their original thinking and for the first time see that understanding of the science concepts—I think that was a good moment. Especially the students with IEPs, it was the first time I could see what they knew about science. [Teacher Comment]

Practical challenges. Without exception, teachers and students commented on the frustration they felt in the practical challenges they faced in the using the UDSN. These challenges were unique to the particular issues within each school. All but two schools in the study had hardware-related challenges, with only four or five computers in each classroom and a computer lab that was sometimes difficult to schedule, “I only had four computers for them to use, and I have 30 children, so that was a barrier” [Teacher Comment].

I had five computers here in my classroom. They used those and then I sent some into the computer lab. I would ask the computer teacher how many computers are free in the lab, and then I would send that number of students down there. And then in the media center, I signed up for blocks of time where we would have 10 computers. So my kids were kind of farmed out all over the building at times! [Teacher Comment]

Most schools lacked access to adequate broadband given the number of students who were trying to access the Internet at the same time (including for classes not using the UDSN) and the amount of data that needed to be transferred. For students and teachers, the result was a sense that the UDSN was not working fast enough, “one bad thing I had myself, was it took too long to load, that was really frustrating” [Student Comment]. Most teachers noted that they had access to some technology but that the technology that they had access to lagged behind the kinds of things they wanted to use the technology for in their teaching,

at least at this school, we're not technology starved, but we still don't have access like we really should. I have ideas about things to do and would like to do with technology. There are lots of interesting programs out there that could be helpful to me, but most of the time I don't bother because it's so hard to deal with the technology in my school [Teacher Comment].

Discussion

Previous research on the use and effectiveness of science notebooks demonstrates the substantial challenge of supporting struggling students to effectively engage in desired science learning behaviors (Englert et al., 1988; Graham, 1990; Graham et al., 1991; Swanson, 1999; Wong, 2001). Yet, the results of this investigation indicate that the UDSN was successful in fostering improved outcomes in science knowledge, as compared with traditional paper-and-pencil science notebooks. The UDSN “raised all boats,” including for those students who exhibited low reading and writing proficiency, and low motivation for science at pretest. These findings are especially remarkable given that students used the UDSN only an average of one time each week, and only an average of 10 times total over the course of the implementation. What aspects of the UDSN contributed to these striking outcomes? The UDSN was designed both to overcome accessibility-related and construct-irrelevant barriers to learning and to provide contextual supports that promote the deep science learning intended through the use of science notebooks. How did these intentions play out in use of the UDSN?

Although for each individual student, we are not able to disentangle which specific aspects of the UDSN's design were most critical, this kind of analysis would not necessarily provide meaningful insight according to the learning design used. Although certain features are necessary for accessibility for certain populations (e.g., alt text on images for students with low vision), the emphasis in design was not on providing particular features for particular audiences, but rather on including options and contextual supports that are likely to improve access to learning for all—this is the UDL approach. A given student might in one notebook session benefit most from using text-to-speech to understand the procedure in an inquiry-based activity, and the same student might later have little difficulty with a reading segment but use an animated coach to think about how to craft an explanation. As suggested by the interviews and the finding of equal benefits across students of varying levels of ability and motivation, this flexibility seemed to have an overall positive impact. We turn now to aspects of that gestalt that emerge as likely contributors to effectiveness of the UDSN.

User-Experience Design and UDL

Technology in and of itself is not a means to successful learning outcomes, but rather a more flexible platform on which content and learning experiences can be rendered. The UDSN leveraged this flexibility to enact the theory of change and design of the program using the principles of UDL. The purpose of UDL is to facilitate the creation and study of learning environments that are usable by and effective for as many learners as possible. It is an approach that attempts to leverage the learning sciences in the user-experience design of educational environments, a kind of

continuously improving framework for active translation between research and practice (Rappolt-Schlichtmann et al., 2012).

The design of education experiences when UDL is leveraged should expressly focus on creating “desirable difficulties” for students, while reducing construct-irrelevant barriers to learning (Bjork, 1994). In this way, designers focus on rendering environments that create challenges for students that are most central to the targeted learning goal or process, while simultaneously reducing the unintended effects of obstacles that are tangential to key learning goals. The resulting experience is one in which students feel competent and confident to share and show what they know.

Reducing construct-irrelevant barriers. This kind of user-experience outcome was evident in the UDSN implementation in which teachers reported seeing students’ original thinking for the first time, and students reported feeling a renewed sense of ownership over and competence with their science work. For example, providing multiple means of expression (a UDL design principle) through the UDSN platform allowed students with disabilities and those with handwriting or expressive difficulty alike to demonstrate their science knowledge to their teacher through means that were most effective for them. Teachers then had a more productive platform to engage in a recursive feedback and revision process with students, targeted to students’ specific level of science understanding.

With paper-and-pencil science notebooks, teachers’ knowledge of students’ understanding of science is more likely to be obscured because student explanations are at once a representation of their competence at written expression and their thinking about science. Although handwriting may be an important learning goal in some instances, there is no need for it to be a barrier to some students in reaping the benefits of a science notebook; in this case, building science content knowledge is the primary learning goal. Importantly, the writing process as it supports science learning through the use of the science notebook is maintained, but those students with low literacy levels can alternatively choose to record data and/or compose explanations by audio recording or drawing. And, contextual supports are provided to remediate other barriers like spelling and handwriting. This tight focus on reducing construct-irrelevant barriers, while amplifying the central goals of the curriculum through the provision of contextual supports in a flexible, digital learning environment, allowed all students to have access to learning and express their thinking about science in ways that were accessible to their teachers.

Creating desirable difficulty. With construct-irrelevant barriers reduced or eliminated, contextual supports can be levied to create “desirable difficulty.” The learning design is purposefully shaped to allow students to calibrate their own levels of challenge, without diluting the science concepts and productive scientific behaviors. Research from various perspectives emphasizes the key role of balance between the level of challenge in the environment and one’s perceived skills and resources as the driving force in shaping affective responses and cognitive engagement (Blascovich, Mendes, Tomaka, Saloman, & Seery, 2003; Csikszentmihalyi, 1991; Daley & Rappolt-Schlichtmann, 2009; Lazarus & Folkman, 1984). For example, Blascovich and colleagues describe “challenge” motivational states when an individual perceives his or her resources as in balance with the demands of a task (Blascovich et al., 2003). Challenge states promote cognitive flexibility and decision making and are characterized by energized, active psycho-

physiological states. In a related framework, Lazarus and Folkman (1984) provide a model of appraisal and adaptation in which positive emotions emerge from “challenging” experiences characterized by closely leveled demands and resources; such experiences lead to the mobilization of energy and promote the effort to respond.

UDL leverages these concepts to provide a framework by which designers within education can systematically consider the provision of contextual supports to create balanced appraisals of learning challenges by diverse students. Such balanced appraisals create the conditions necessary for deep engagement to occur. Engagement in learning is at once emotional and cognitive, and is achieved through the application of appropriate challenge and calibrated to individual learners’ specific strengths and weaknesses. A student can choose to access or ignore a given support, to use any of the various means of responding to a prompt, and to watch or pass by a video that provides additional information. Designers anticipate and reduce or eliminate barriers to deep engagement by providing options and supports that render the learning environment flexible and maintain a focus on specific learning goals (Meyer & Rose, 1998; Rose & Meyer, 2002). Technology enhances the degree of flexibility.

The importance of contextual supports to the productive use of the UDSN and the creation of engagement in science learning was confirmed in both the quantitative and qualitative analyses. Teachers commented on the utility of the contextual supports for students in their science learning, and even hypothesized as to the emotional benefits in reducing anxiety and promoting more independent, confident work. When prompted to think about and share how the UDSN did or did not help their science learning, students often expressed feelings of agency and confidence knowing that the UDSN offered resources to help them if needed. Some indicated that when using supports, building explanations seemed more “doable,” where success was possible even though the task felt difficult.

In the qualitative research, competence (Harter, 1978; R. W. White, 1963) and autonomy (deCharms, 1968; Deci, 1975) surfaced as key themes among students and teachers in reporting their perceptions of the usefulness of the UDSN to science learning. Although understandable, this finding was not expected and suggests an avenue for future work. Students and teachers reported that in using the UDSN, students felt more ownership, agency, and control over the work as they were supported to build skills and feel competent in the executing of the inquiry process and producing explanations describing their science experiences. It may be that the overwhelmingly high and persistent levels of interest and excitement reported among students using the UDSN were, at least in part, attributable to the generation of feelings of competence and autonomy in their work. Building from self-determination theory (SDT), where educational contexts are seen to catalyze within and between person differences in motivation (Deci & Ryan, 1985, 1991; R. M. Ryan, 1995), a stronger focus on creating a sense of relatedness (Baumeister & Leary, 1995; Reis, 1994) through the design of the UDSN may have further optimized student’s performance, engagement, and feelings of well-being. It would be useful to do a deeper analysis relating the engagement principle from the UDL framework with the research literature on SDT especially with regard to those research practice models that experimentally

describe conditions that foster versus undermine human potential in learning.

Designing for Variability in Engagement and Learning

It is important to note that although most students thought that contextual supports were useful to their science learning, students were highly variable in whether they typically used supports and found them helpful. This was not unexpected. Taking a cue from the learning sciences, UDL assumes that variability is the rule and not the exception in learning and that learning is actively organized and context specific (Fischer & Bidell, 2006; Plomin & Kovas, 2005; Rappolt-Schlichtmann, Tenenbaum, Koepke, & Fischer, 2007; Thelen & Smith, 1994; Van Geert, 1998). Learners differ markedly in the ways in which they can be engaged or motivated to learn, so when designing contextual supports using the UDL framework, developers construct environments that offer multiple means and options to provide access to supports that should allow students to reappraise challenging tasks as demanding but doable.

One major challenge that emerges from this approach is that students vary in the degree to which they make active and good choices about the supports they leverage, and in fact the literature suggests that students who would most benefit from embedded supports are often least likely to choose to access them (A. M. Ryan & Pintrich, 1997, 1998; A. M. Ryan, Pintrich, & Midgley, 2001). For this reason, teachers play an important and specific role in the implementation of support-rich UDL environments. They are facilitators and mediators, helping students to learn how to best leverage the designed environment in the service of learning. They are keen observers and effective users of data to understand the strengths and weaknesses of their students and, when necessary, guide the relationship between the student and the learning environment.

Our quantitative findings reflect these challenges. The guidance teachers offered students in leveraging the UDSN and the approach to instruction teachers used clearly played a role in how students used the UDSN. In the context of UDL-designed environments, instruction that is (a) adaptable to student strengths and weaknesses as students change and (b) carefully planned but responsive to “in the moment” teaching opportunities will be more effective than one-size-fits all, static approaches (Connor et al., 2009). Additionally, the effective use of real-time opportunities depends on teachers’ levels of expertise with the pedagogical approach, the curriculum, the learning goals, and the students’ needs. Indeed, we found that students whose teachers had more experience using science notebooks in instruction tended to demonstrate higher frequency of process behaviors.

Limitations

With regard to the design of the research study, we were not able to consider what process behaviors were used in the control group. Without an equivalent of the electronic usage log, we could not determine how often, for example, students reviewed data when entering explanations or how often they revised previous notebook entries. Future work could incorporate this type of comparison.

In addition and while the impact of the UDSN was positive in this implementation, teachers generally used the UDSN much less frequently than intended. The web-based notebook was meant to

be a part of regular science activities—a place to record data during observations, take notes during interactions with teachers, and work regularly throughout the curriculum. Instead, it often became a place for weekly entering of previously collected data and focused reflection. In part, this limitation reflects the state of broadband and computer access in public schools.

Access to adequate hardware and the Internet is still a significant problem in American public schools. Estimates indicate that 80% of public schools do not have Internet access adequate for their instructional needs, and school leadership at the elementary level often underestimates the degree to which teachers and students can use computers and especially the Internet as a part of the normal course of teaching and learning (Fox, Waters, Fletcher, & Levin, 2012). Studies like this one should help to raise awareness as to the utility of elementary level computer and Internet use, as well as the limitations of current infrastructure.

Conclusion

Interest in the UDL framework has increased exponentially over the last decade. The Higher Education Opportunity Act of 2008 (HEOA, 2012, 20 U.S.C. § 1003(24)) established the statutory definition for UDL, strongly suggesting that preservice teacher training incorporate instruction on strategies consistent with UDL (HEOA, 2012, 20 U.S.C. § 1022d(b)(1)(K)), and the U.S. Department of Education’s National Educational Technology Plan 2010 makes frequent reference to UDL as a framework that reduces barriers and maximizes learning opportunities for all students (U.S. Department of Education, Office of Educational Technology, 2010). This interest is understandable because the framework provides a possible answer to the growing call for more “personalized” curricular materials that have the potential to accommodate the full diversity of learners and teachers within the education system, and, furthermore, the UDL framework explicitly reflects our best understanding of the learning sciences as it can be applied to education design. However, research exploring UDL as an approach to education design is still in its infancy.

This study is one of only a handful of experimental studies looking at either the overall impact of a UDL technology in authentic classrooms or the active components of UDL technology in the process of development. To our knowledge, this is the only study to qualitatively explore students’ and teachers’ experiences and perceptions of usefulness of UDL technology to learning. There is substantial opportunity to explore, examine, and inform theory and research concerning the nature of learning and development from a practice-oriented perspective. Such work along with implementation and design-based research on UDL environments is warranted. Defining a research agenda for UDL is beyond the scope of this article, but several other published articles make suggestions in this regard (e.g., Rappolt-Schlichtmann et al., 2012).

Digital technologies used in conjunction with strong teaching strategies offer unprecedented opportunities to support developing active science learning skills, but the technology-based format does not automatically improve on the print format. Many technology-based programs and digital materials are inaccessible, and the impact of existing technology-based programs, even when effective, are typically small (e.g., see “IES What Works Clearinghouse”; <http://ies.ed.gov/ncee/wwc/reports/Topicarea.aspx>).

tid = 15). However, as this work demonstrates, when technology is used to foster a supported learning environment in which the emphasis is on core learning activities, with strong teacher experience and embedded support for construct-irrelevant skills and strategies, technology can provide consistent gains for a variety of learners.

References

- Banilower, E. R. (2002). *2001–2002 study of the impact of LSC initiative on student achievement in science*. Chapel Hill, NC: Horizon Research, Inc.
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117, 497–529. doi:10.1037/0033-2909.117.3.497
- Baxter, G. P., Bass, K. M., & Glaser, R. (2001). Notebook writing in three fifth-grade science classrooms. *Elementary School Journal*, 102, 123–140. doi:10.1086/499696
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.) *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.
- Blascovich, J., Mendes, W. B., Tomaka, J., Salomon, K., & Seery, M. D. (2003). The robust nature of the biopsychosocial model of challenge and threat: A reply to Wright and Kirby. *Personality and Social Psychology Review*, 7, 234–243. doi:10.1207/S15327957PSPR0703_03
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65, 245–281. doi:10.3102/00346543065003245
- CAST. (2011). *Universal design for learning guidelines version 2.0*. Retrieved from <http://www.udlcenter.org/aboutudl/udlguidelines/downloads>
- Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, 86, 175–218. doi:10.1002/sce.10001
- Cobb, P. (2001). Supporting the improvement of learning and teaching in social and institutional context. In S. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 455–478). Cambridge, MA: Lawrence Erlbaum Associates.
- Cognition and Technology Group at Vanderbilt. (1993). Examining the cognitive challenges and pedagogical opportunities of integrated media systems: Toward a research agenda. *Journal of Special Education Technology*, 12, 118–124.
- Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O'Shea (Eds.), *New directions in educational technology* (pp. 15–22). New York, NY: Springer-Verlag. doi:10.1007/978-3-642-77750-9_2
- Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser* (pp. 453–494). Hillsdale, NJ: Erlbaum.
- Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *Journal of the Learning Sciences*, 13, 15–42. doi:10.1207/s15327809jls1301_2
- Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., Underwood, P., & Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of Child \times Instruction interactions on first graders' literacy development. *Child Development*, 80, 77–100. doi:10.1111/j.1467-8624.2008.01247.x
- Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage.
- Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience*. New York, NY: Harper & Row.
- Daley, S. G., & Rappolt-Schlichtmann, G. (2009, May). *The educational relevance of stress physiology*. Poster presented at the International Mind, Brain, and Education Society National Conference, Philadelphia, PA.
- Dalton, B., Pisha, B., Eagleton, M., Coyne, P., & Deysher, S. (2002). *Engaging the text: Reciprocal teaching and questioning strategies in a scaffolded learning environment. Final report to the U.S. Department of Education*. Peabody, MA: CAST.
- deCharms, R. (1968). *Personal causation*. New York, NY: Academic Press.
- Deci, E. L. (1975). *Intrinsic motivation*. New York, NY: Plenum Press. doi:10.1007/978-1-4613-4446-9
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York, NY: Plenum Press.
- Deci, E. L., & Ryan, R. M. (1991). A motivational approach to self: Integration in personality. In R. Dienstbier (Ed.), *Nebraska Symposium on Motivation: Vol. 38. Perspectives on motivation* (pp. 237–288). Lincoln: University of Nebraska Press.
- De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, 68, 179–201. doi:10.3102/00346543068002179
- Dolan, R. P., Hall, T. E., Banerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment*, 3, 4–31.
- Englert, C., Raphael, T., Fear, K., & Anderson, L. (1988). Students' metacognitive knowledge about how to write informational texts. *Learning Disability Quarterly*, 11, 18–46. doi:10.2307/1511035
- Ferguson, G., Long, K., & Kennedy, C. (2009, September). *Assessing science knowledge: Implementation through teacher research (ASK-IT Final Report)*. Retrieved May 24, 2010, from http://www.wastatelaser.org/_news/images/ASK-IT_Final-report_09032009.pdf
- Fischer, K. W., & Bidell, T. R. (2006). Dynamic development of action, thought, and emotion. In R. M. Lerner (Ed.), *Theoretical models of human development* (6th ed., pp. 313–399). New York, NY: Wiley.
- Flagg, B. N. (1990). *Formative evaluation for educational technologies*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fox, C., Waters, J., Fletcher, G., & Levin, D. (2012). *The broadband imperative: Recommendations to address K-12 education infrastructure needs*. Washington, DC: State Educational Technology Directors Association (SETDA).
- Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine Publishing Company.
- Graham, S. (1990). The role of production factors in learning disabled students compositions. *Journal of Educational Psychology*, 82, 781–791. doi:10.1037/0022-0663.82.4.781
- Graham, S., Harris, K., MacArthur, C., & Schwartz, S. (1991). Writing and writing instruction with students with learning disabilities: A review of a program of research. *Learning Disability Quarterly*, 14, 89–114. doi:10.2307/1510517
- Hargrove, T. Y., & Nesbit, C. (2003). *Science notebooks: Tools for increasing achievement across the curriculum* (ERIC Document Reproduction Service Number ED 482720). Columbus, OH: ERIC Clearinghouse for Science Mathematics and Environmental Education.
- Harter, S. (1978). Effectance motivation reconsidered: Toward a developmental model. *Human Development*, 21, 34–64.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77, 81–112. doi:10.3102/003465430298487
- Higher Education Opportunity Act of 2008, 20 U.S.C. § 1001 et seq. (2012).
- Hsu, Y. S. (2004). Using the Internet to develop students' capacity for scientific inquiry. *Journal of Educational Computing Research*, 31, 137–161. doi:10.2190/HYX8-CK1A-FVU3-5Y5W

- Individuals with Disabilities Education Act of 2004, 20 U.S.C. § 1400 et seq.
- Kame'enui, E. J., & Carnine, D. W. (1998). *Effective teaching strategies that accommodate diverse learners*. Columbus, OH: Merrill.
- Keselman, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *Journal of Research in Science Teaching*, 40, 898–921. doi:10.1002/tea.10115
- Keys, C. W. (2000). Investigating the thinking processes of eighth grade writers during the composition of a scientific laboratory report. *Journal of Research in Science Teaching*, 37, 676–690. doi:10.1002/1098-2736(200009)37:7<676::AID-TEA4>3.0.CO;2-6
- Klecan-Aker, J. S., & Caraway, T. H. (1997). A study of the relationship of storytelling ability and reading comprehension in fourth and sixth grade African-American children. *International Journal of Language & Communication Disorders*, 32, 109–125. doi:10.3109/13682829709021464
- Klentschy, M. (2005). Science notebook essentials. *Science and Children*, 43, 24–27.
- Klentschy, M., Garrison, L., & Amaral, O. M. (1999). *Valle Imperial Project in Science (VIPs): Four-year comparison of student achievement data 1995–1999*. Retrieved from www.lawrencehallofscience.org/foss/scope/research/VIPStudy.pdf
- Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. New York, NY: Springer.
- LeCompte, M. D., & Preissle, J. (1992). Toward an ethology of student life in classrooms: Synthesizing the qualitative research tradition. In M. D. LeCompte, W. Millroy, & J. Preissle (Eds.), *The handbook of qualitative research in education* (pp. 815–861). San Diego, CA: Academic Press.
- Mastropieri, M. A., Scruggs, T. E., Boon, R., & Carter, K. B. (2001). Correlates of inquiry learning in science: Constructing concepts of density and buoyancy. *Remedial and Special Education*, 22, 130–137. doi:10.1177/074193250102200301
- McNeill, K., & Krajcik, J. (2006, April). *Supporting students' constructions of scientific explanation through generic versus context-specific written scaffolds*. Paper presented at the AERA annual meeting, San Francisco, CA.
- Meyer, A., & Rose, D. (1998). Learning to read in the computer age. In J. Chall (Series Ed.), J. Onofrey (Ed.), *Reading research to practice* (pp. xxx–xxx). Cambridge, MA: Brookline Books.
- Miles, M., & Hubberman, A. (1984). *Qualitative data analysis: A source book of new methods*. Beverly Hills, CA: Sage.
- Miller, R. G., & Calfee, R. C. (2004). Making thinking visible. *Science and Children*, 42, 20–25.
- National Center for Accessible Media. (2006). *Making educational software and web sites accessible: Design guidelines including math and science solutions*. Retrieved from http://ncam.wgbh.org/invent_build/web_multimedia/accessible-digital-media-guide/
- National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.
- National Research Council. (2000). *Inquiry and the national science education standards*. Washington, DC: National Academy Press.
- No Child Left Behind Act of 2001, 20 U.S.C. § 6301 et seq.
- Northwest Evaluation Association. (2005). *RIT scale norms for use with achievement level tests and Measures of Academic Progress*. Lake Oswego, OR: Author.
- NUD*IST nVIVO. (2002). nVIVO (Version 7). QSR International [Computer software]. Doncaster, Victoria Australia. Retrieved from www.qrsinternational.com
- Palinscar, A. S. (1986). Metacognitive strategy instruction. *Exceptional Children*, 53, 118–124.
- Palinscar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review of Psychology*, 49, 345–375. doi:10.1146/annurev.psych.49.1.345
- Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1, 117–175. doi:10.1207/s1532690xci0102_1
- Plomin, R., & Kovas, Y. (2005). Generalist genes and learning disabilities. *Psychological Bulletin*, 131, 592–617. doi:10.1037/0033-2909.131.4.592
- Rappolt-Schlichtmann, G., Daley, S., Lim, S., Robinson, K., & Johnson, M. (2011). *The universally designed science notebook: An intervention to support science learning for students with disabilities. Final Report submitted to the U.S. Department of Education, Institute of Education Sciences*. Washington, DC: CAST.
- Rappolt-Schlichtmann, G., Daley, S. G., & Rose, L. T. (Eds.). (2012). *A research reader in universal design for learning*. Cambridge, MA: Harvard Education Press.
- Rappolt-Schlichtmann, G., Tenenbaum, H., Koepke, M., & Fischer, K. (2007). Transient and robust knowledge: Contextual support and the dynamics of children's reasoning about density. *Mind, Brain, and Education*, 1, 98–108. doi:10.1111/j.1751-228X.2007.00010.x
- Rappolt-Schlichtmann, G., & Watamura, S. (2010). Inter- and transdisciplinary work: Connecting research on hormones with problems of educational practice. *Mind, Brain, and Education*, 4, 156–158. doi:10.1111/j.1751-228X.2010.01094.x
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Reeves, T. C., & Hedberg, J. G. (2003). *Interactive learning systems evaluation*. Englewood Cliffs, NJ: Educational Technology Publications.
- Rehabilitation Act, Section 508 of 29 U.S.C. § 794d.
- Reis, H. T. (1994). Domains of experience: Investigating relationship processes from three perspectives. In R. Erber & R. Gilmour (Eds.), *Theoretical frameworks for personal relationships* (pp. 87–110). Hillsdale, NJ: Erlbaum.
- Rose, D. H., & Meyer, A. (2002). *Teaching every student in the digital age: Universal design for learning*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Rose, D. H., & Meyer, A. (Eds.). (2006). *A practical reader in Universal Design for Learning*. Cambridge, MA: Harvard Education Press.
- Ruiz-Primo, M., Li, M., Ayala, C., & Shavelson, R. (2004). Evaluating students' science notebooks as an assessment tool: Research report. *International Journal of Science Education*, 26, 1477–1506. doi:10.1080/0950069042000177299
- Ryan, A. M., & Pintrich, P. R. (1997). "Should I ask for help?" The role of motivation and attitudes in adolescents' help seeking in math class. *Journal of Educational Psychology*, 89, 329–341. doi:10.1037/0022-0663.89.2.329
- Ryan, A. M., & Pintrich, P. R. (1998). Achievement and social motivational influences on help-seeking in the classroom. In S. A. Karabenick (Ed.), *Strategic help-seeking: Implications for learning and teaching* (pp. 117–139). Mahwah, NJ: Erlbaum.
- Ryan, A. M., Pintrich, P. R., & Midgley, C. (2001). Avoiding seeking help in the classroom: Who and why? *Educational Psychology Review*, 13, 93–114. doi:10.1023/A:1009013420053
- Ryan, R. M. (1995). Psychological needs and the facilitation of integrative processes. *Journal of Personality*, 63, 397–427. doi:10.1111/j.1467-6494.1995.tb00501.x
- Samuels, C. (2009). Universal design concept pushed for in education. In D. T. Gordon, J. W. Gravel, & L. A. Schifter (Eds.), *A policy reader in universal design for learning* (pp. 35–45). Cambridge, MA: Harvard Education Press.
- Sandoval, W., & Reiser, B. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, 88, 345–372. doi:10.1002/sce.10130

- Scruggs, T. E., & Mastropieri, M. A. (1994). The construction of scientific knowledge by students with mild disabilities. *Journal of Special Education*, 28, 307–321. doi:10.1177/002246699402800306
- Scruggs, T. E., Mastropieri, M. A., Bakken, J. P., & Brigham, F. J. (1993). Reading versus doing: The relative effects of textbook-based and inquiry-oriented approaches to science learning in special education classrooms. *Journal of Special Education*, 27, 1–15. doi:10.1177/002246699302700101
- Shepardson, D. P., & Britsch, S. J. (2004). The art of reviewing science journals. *Science and Children*, 42, 43–45.
- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 23, 323–355.
- Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology*, 38, 934–947. doi:10.1037/0012-1649.38.6.934
- Swanson, H. L. (1999). Reading comprehension and working memory in learning-disabled readers: Is the phonological loop more important than the executive system? *Journal of Experimental Child Psychology*, 72, 1–31. doi:10.1006/jecp.1998.2477
- Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

- U.S. Department of Education, Office of Educational Technology. (2010). *Transforming American education: Learning powered by technology* (National Education Technology Plan 2010). Washington, DC: Author. Retrieved from <http://www.ed.gov/sites/default/files/netp2010.pdf>
- Van Geert, P. (1998). A dynamic systems model of basic developmental mechanisms: Piaget, Vygotsky, and beyond. *Psychological Review*, 105, 5, 634–677.
- White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and meta-cognition: Making science accessible to all students. *Cognition and Instruction*, 16, 3–118. doi:10.1207/s1532690xci1601_2
- White, R. W. (1963). *Ego and reality in psychoanalytic theory*. New York, NY: International Universities Press.
- Wong, B. Y. L. (2001). Commentary: Pointers for literacy instruction from educational technology and research on writing instruction. *The Elementary School Journal*, 101, 359–369. doi:10.1086/499674
- World Wide Web Consortium Web Accessibility Initiative. (1999). Web content accessibility guidelines 1.0. Retrieved from <http://www.w3.org/TR/1999/WAI-WEBCONTENT-19990505/>

Received December 15, 2011

Revision received April 1, 2013

Accepted April 5, 2013 ■

UNITED STATES POSTAL SERVICE® (All Periodicals Publications Except Requester Publications)

Statement of Ownership, Management, and Circulation

1. Publication Title: Journal of Educational Psychology

2. Issue Frequency: Quarterly

3. Issue Date for Circulation Data Below: October 2013

4. Annual Subscription Price: Indiv \$100

5. Number of Issues Published Annually: 4

6. Annual Subscription Price: Indiv \$100

7. Complete Mailing Address of Known Office of Publication (Not printer): 750 First Street, NE, Washington, DC 20002-4242

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer): 750 First Street, NE, Washington, DC 20002-4242

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank):

Publisher: American Psychological Association
750 First Street, NE
Washington, DC 20002-4242
 Editor: Art Wapner, PhD, Department of Psychology
University of Georgia, 172 Psychology Building
Athens, GA 30602-3030
 Managing Editor: Jack Horvitz
American Psychological Association
750 First Street, NE, Washington, DC 20002-4242

10. Complete Mailing Address of the Circulation Office (Do not leave blank):

Full Name: American Psychological Association
 Complete Mailing Address: 750 First Street, NE
Washington, DC 20002-4242

11. For completion by nonprofit organizations authorized to mail at nonprofit rates (Check one):

Publication Title: Journal of Educational Psychology

12. Tax Status (For completion by nonprofit organizations authorized to mail at nonprofit rates (Check one):

Publication Title: Journal of Educational Psychology

13. Publication Title: Journal of Educational Psychology

14. Publication Title: Journal of Educational Psychology

15. Publication Title: Journal of Educational Psychology

16. Publication Title: Journal of Educational Psychology

17. Publication Title: Journal of Educational Psychology

18. Publication Title: Journal of Educational Psychology

19. Publication Title: Journal of Educational Psychology

20. Publication Title: Journal of Educational Psychology

21. Publication Title: Journal of Educational Psychology

22. Publication Title: Journal of Educational Psychology

23. Publication Title: Journal of Educational Psychology

24. Publication Title: Journal of Educational Psychology

25. Publication Title: Journal of Educational Psychology

26. Publication Title: Journal of Educational Psychology

27. Publication Title: Journal of Educational Psychology

28. Publication Title: Journal of Educational Psychology

29. Publication Title: Journal of Educational Psychology

30. Publication Title: Journal of Educational Psychology

31. Publication Title: Journal of Educational Psychology

32. Publication Title: Journal of Educational Psychology

33. Publication Title: Journal of Educational Psychology

34. Publication Title: Journal of Educational Psychology

35. Publication Title: Journal of Educational Psychology

36. Publication Title: Journal of Educational Psychology

37. Publication Title: Journal of Educational Psychology

38. Publication Title: Journal of Educational Psychology

39. Publication Title: Journal of Educational Psychology

40. Publication Title: Journal of Educational Psychology

41. Publication Title: Journal of Educational Psychology

42. Publication Title: Journal of Educational Psychology

43. Publication Title: Journal of Educational Psychology

44. Publication Title: Journal of Educational Psychology

45. Publication Title: Journal of Educational Psychology

46. Publication Title: Journal of Educational Psychology

47. Publication Title: Journal of Educational Psychology

48. Publication Title: Journal of Educational Psychology

49. Publication Title: Journal of Educational Psychology

50. Publication Title: Journal of Educational Psychology

51. Publication Title: Journal of Educational Psychology

52. Publication Title: Journal of Educational Psychology

53. Publication Title: Journal of Educational Psychology

54. Publication Title: Journal of Educational Psychology

55. Publication Title: Journal of Educational Psychology

56. Publication Title: Journal of Educational Psychology

57. Publication Title: Journal of Educational Psychology

58. Publication Title: Journal of Educational Psychology

59. Publication Title: Journal of Educational Psychology

60. Publication Title: Journal of Educational Psychology

61. Publication Title: Journal of Educational Psychology

62. Publication Title: Journal of Educational Psychology

63. Publication Title: Journal of Educational Psychology

64. Publication Title: Journal of Educational Psychology

65. Publication Title: Journal of Educational Psychology

66. Publication Title: Journal of Educational Psychology

67. Publication Title: Journal of Educational Psychology

68. Publication Title: Journal of Educational Psychology

69. Publication Title: Journal of Educational Psychology

70. Publication Title: Journal of Educational Psychology

71. Publication Title: Journal of Educational Psychology

72. Publication Title: Journal of Educational Psychology

73. Publication Title: Journal of Educational Psychology

74. Publication Title: Journal of Educational Psychology

75. Publication Title: Journal of Educational Psychology

76. Publication Title: Journal of Educational Psychology

77. Publication Title: Journal of Educational Psychology

78. Publication Title: Journal of Educational Psychology

79. Publication Title: Journal of Educational Psychology

80. Publication Title: Journal of Educational Psychology

81. Publication Title: Journal of Educational Psychology

82. Publication Title: Journal of Educational Psychology

83. Publication Title: Journal of Educational Psychology

84. Publication Title: Journal of Educational Psychology

85. Publication Title: Journal of Educational Psychology

86. Publication Title: Journal of Educational Psychology

87. Publication Title: Journal of Educational Psychology

88. Publication Title: Journal of Educational Psychology

89. Publication Title: Journal of Educational Psychology

90. Publication Title: Journal of Educational Psychology

91. Publication Title: Journal of Educational Psychology

92. Publication Title: Journal of Educational Psychology

93. Publication Title: Journal of Educational Psychology

94. Publication Title: Journal of Educational Psychology

95. Publication Title: Journal of Educational Psychology

96. Publication Title: Journal of Educational Psychology

97. Publication Title: Journal of Educational Psychology

98. Publication Title: Journal of Educational Psychology

99. Publication Title: Journal of Educational Psychology

100. Publication Title: Journal of Educational Psychology

13. Publication Title: Journal of Educational Psychology

14. Issue Date for Circulation Data Below: August 2013

15. Extent and Nature of Circulation

15. Extent and Nature of Circulation	Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net paid and unpaid)	1713	1709
b. Paid Circulation (Net paid and unpaid)	1139	1108
c. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
d. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
e. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
f. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
g. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
h. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
i. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
j. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
k. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
l. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
m. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
n. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
o. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
p. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
q. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
r. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
s. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
t. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
u. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
v. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
w. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
x. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
y. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
z. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
aa. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ab. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ac. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ad. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ae. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
af. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ag. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ah. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ai. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
aj. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ak. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
al. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
am. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
an. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ao. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ap. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
aq. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ar. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
as. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
at. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
au. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
av. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
aw. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ax. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ay. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
az. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ba. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bb. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bc. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bd. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
be. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bf. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bg. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bh. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bi. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bj. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bk. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bl. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bm. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bn. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bo. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bp. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bq. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
br. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bs. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bt. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bu. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bv. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bw. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bx. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
by. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
bz. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ca. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cb. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cc. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cd. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ce. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cf. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cg. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ch. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ci. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cj. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ck. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cl. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cm. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cn. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
co. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cp. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cq. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cr. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cs. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ct. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cu. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cv. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cw. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cx. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cy. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
cz. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
da. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
db. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dc. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dd. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
de. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
df. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dg. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dh. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
di. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dj. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dk. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dl. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dm. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dn. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
do. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dp. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dq. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dr. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ds. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dt. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
du. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dv. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dw. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dx. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dy. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
dz. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 15b(3))	1514	1484
ea. Free or Nominal Rate (Sum of 15b(1), 15b(2), and 1		

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Manuscript preparation. Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see www.apa.org/pubs/journals/edu. **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

- Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin*, 139, 133–151. <http://dx.doi.org/10.1037/a0028566>
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.
- Gill, M. J., & Sypher, B. D. (2009). Workplace incivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied as Tiff, EPS, or PowerPoint. The minimum line weight for line art is 0.5 point for optimal printing. Original color figures can be printed in color at the editor's and publisher's discretion provided the author agrees to pay \$255 for one figure, \$425 for two figures, \$575 for three figures, \$675 for four figures, and \$55 for each additional figure.

Publication policies. APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at www.apa.org/pubs/authors/posting.aspx. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult

journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Masked review policy. The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . ." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . ." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at www.apa.org/ethics/ or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

Permissions. Authors of accepted papers are required to obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including, for example, test materials or portions thereof and photographs of people.

Supplemental materials. APA can now place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see www.apa.org/pubs/authors/supp-material.aspx for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

Submission. Authors should submit their manuscripts electronically via the Manuscript Submission Portal at www.apa.org/pubs/journals/edu/index.aspx (follow the link for submission under Instructions to Authors). General correspondence may be addressed to the editorial office at jedgar@memphis.edu.

Preparing files for production. If your manuscript is accepted for publication, please follow the guidelines for file formats and naming provided at www.apa.org/pubs/journals/authors/preparing-efiles.aspx. If your manuscript was mask reviewed, please ensure that the final version for production includes a byline and full author note for typesetting.

Acknowledgments

The editor thanks the following principal reviewers who evaluated at least 3 manuscripts for *Journal of Educational Psychology* between June 1, 2012 and May 31, 2013.

Matthew L. Bernacki	Annemarie Hindman	Andrew J. Martin	Greg Roberts
Catherine Bohn-Gettler	Scott R. Hinze	Margaret G. McKeown	Doug Rohrer
Jason L. G. Braasch	Flaviu A. Hodis	Kou Murayama	Linda Rose-Krasnor
Matthew Burns			
Andrew Butler	Eric D. Jones	Nikos Ntoumaris	
	Uta Klusman	Paul A. O'Keefe	Wolfgang Schnotz
Simona C. S. Caravita	Evenlyn Kroesbergen		Sungok Serena Shim
		Sharolyn D. Pollard-Durodola	Seung-Hee Son
David K. Dickinson		David J. Purpura	Rayne A. Sperling
	Jonas W. B. Lang		Joseph J. Stevens
Jim Fryer	Che Kan Leong	Gerardo Ramirez	H. Lee Swanson
	Detley Leutner	Robert Reid	
Michael Graves	Gregory Arief Liem	Richard Remedios	Holly A. Taylor
Jeffrey A. Greene	David Lubinski	Lindsey Richland	Rebecca Treiman
Frederic Guay			

The editor also thanks the following ad hoc reviewers who evaluated manuscripts for *Journal of Educational Psychology* between June 1, 2012 and May 31, 2013.

Amy Adcock	Brian Bottge	Samuel Day	David Geary
Jill L. Adelson	Robert H. Bradley	Stephanie Day	Peter Gerjets
Olusola O. Adesope	Catherine Bradshaw	Bieke de Fraine	Wim Gijssels
Wondimu Ahmed	Sarah K. Brem	Peter F. de Jong	Michele Gill
Stephanie Al Otaiba	Rainer Bromme	Bert de Smedt	Robyn M. Gillies
Louis Alfieri	Suzanne H. Broughton	Pascal Deboeck	Inga Glogger
Janice Field Almasi	Tad Brunye	Paul Deboeck	Janice D. Gobert
Steve Amendum	Rebecca Bull	Krista Deleeuw	Thomas Goetz
Jason L. Anthony	Jeni L. Burnette	Andreas Demetriou	Imani Masters Goffney
Kenn Apel		Bert DeSmedt	Claude Goldenberg
Alison W. Arrow	Kate Cain	Irene-Anna N. Diakidoy	Frank Goldhammer
Mark Ashcraft	Claire E. Cameron	Chantelle J. Dowsett	Ilya Goldin
Avi Assor	Hugh W. Catts	Roni Jo Draper	Kathy E. Green
Kaisa Aunola	Raquel Cerdan	Markus Dresel	Daphne Greenberg
Paul Ayres	Marilyn J. Chambliss	Fiona Duff	Samuel Greiff
	Jason Chen	Greg Duncan	Patrick Griffin
Drew H. Bailey	Xi Chen-Bumgardner	Rob Duncan	Marissa Griggs
Linda Baker	Clark Chinn	C. Emily Durbin	Wendy S. Grolnick
Ryan Baker	Mei-Shiu Chiu	Joe Durlak	Daniel Gucciardi
Arnold B. Bakker	Namok Choi	Mark Dynarski	Serfio Guglielmi
Art Baroody	Paul Cirino		Janine Gut
Christopher D. Barr	Amy Claessens	Stephen Ellenbogen	
Roderick W. Barron	Don Compton	Nicole Else-Quest	Zach Hambrick
Hideko Hamada Bassett	F. Corapci	Cynthia A. Erdley	Gillian Hampden-Thompson
Michael D. Beck	Pierre Cormier	Howard T. Everson	Brenda Jones Harden
Michael Becker	Jeffrey Corneilus-White		Peter Hastings
Sybilla Beckmann	Sarah Critten	Thomas Farmer	Jarkko Hautamaki
Avi Ben-Zeev	Steven Crooks	Emilio Ferrer	Neil Heffernan
Aprile D. Benner	Beno Csapo	Jeremy D. Finn	Angie Heine
David B. Berch	Steven Culpepper	Michael T. Ford	Anne Helsdingen
Elizabeth Bernhardt	James Cummins	Barry J. Fraser	Patricio G. Herbst
Virginia Berninger	Timothy W. Curby	Harald Freudenthaler	Jonathan C. Hilpert
Gina Biancarosa		Joachim Funke	Melissa Holt
Rebecca S. Bigler	David Yun Dai		J. Yang Hong
Kathy S. Binder	Celine Darnon	Iddo Gal	William S. Horton
Marcela Borge	Robert Davies	Mirta Galesic	Elizabeth Howard
Sandra Leanne Bosacki	Heather Davis	Patricia Gandara	Wei-Chen Hung
Christy Kim Boscardin	Pamela Davis-Kean	Colleen Ganley	Carol S. Huntsinger

- Tanner Jackson
 Jeremy Jamieson
 Asha K. Jitendra
 Eric D. Jones
 Martin H. Jones
 Paul R. Jones
 Nancy C. Jordan
 Jaana Juvonen

 Sean H.K. Kang
 Susan Kapitanoff
 Allison Kelaher-Young
 Harrison J. Kell
 Melanie M. Keller
 Panayiota Kendeou
 Fazel Keshtkar
 Ursula Kessels
 Michael Kieffer
 Ulf Kieschke
 Kenneth Kiewra
 Yanghee Kim
 Thomas Kindermann
 John R. Kirby
 Rinke Klein Entink
 Jennifer L. Kobrin
 Kenneth R. Koedinger
 Olaf Koeller
 Svjetlana Kolic-Vehovec
 Tuire K. Koponen
 Nate Kornell
 Aaron Kozbelt
 Bracha Kramarski
 Stephan Kroener
 Evelyn Kroesbergen
 Mareike Kunter
 Christopher A. Kurby
 Hans Kuyper

 David Landy
 Denise Larsen
 Thibaud Latour
 Kimberly A. Lawless
 Joshua F. Lawrence
 Jihyun Lee
 Jo-Anne LeFevre
 James D. Lehman
 Juhani Lehto
 Pui Wa Lei
 Erica S. Lembke
 Wolfgang Lenhard
 Che Kan Leong
 Chantal Levesque-Bristol
 Hongli Li
 Phil D. Liu
 Min Liu
 David F. Lohman
 Christopher Lonigan
 Alexandra Loukas
 Patricia A. Lowe
 Shulan Lu
 Wen Luo

 Lars-Erik Malmberg
 Erin A. Maloney
 Jeannette Mancilla-Martinez
 Gwen C. Marchand
 Jon Margerum-Leys
 Herbert W. Marsh
 Rhonda Martinussen
 Jennifer Matjasko
 Camillia Matuk
 Megan McClelland
 Matt McCrudden
 Nele McElvany
 Lyle McKinney
 Kristen McMaster
 Mike Mensink
 Bonnie J. F. Meyer
 Michael Middleton
 David Miele
 Amori Y. Mikami
 Gloria E. Miller
 Raymond B. Miller
 Angela Miller
 Jacob Mishook
 Kristin L. Moilanen
 Kouider Mokhtari
 Suzanne E. Mol
 Jens Moller
 Chris Mueller
 Katharina Mueller
 Christopher Murray

 Jennifer W. Neal
 Ross Nehm
 Sabina Neugebauer
 Nora Newcombe
 Kristie J. Newton
 Florrie Ng
 Gillian M. Nichols
 Markku Niemivirta
 Timothy J. Nokes-Malach

 Amy Ogan
 Alandeom W. Oliveira
 T. C. Oshima

 Sebastien Pacton
 Aaron M. Pallas
 John Pani
 Robert H. Prada
 Scott Paris
 Gregory Park
 Elise Pas
 Philip I. Pavlik
 Bruce Pennington
 Charles Perfetti
 Zoi Apostolia-Philppakos
 Beth Phillips
 Eva Pomerantz
 Paul Poteat
 Sarah R. Powell
 Kristopher J. Preacher

 Franzis Preckel

 Remi Radel
 Geetha Ramani
 Kevin L. Rand
 Catherine Ratelle
 Stephen W. Raudenbush
 Diana Raufelder
 Rachel Razza
 Johnmarshall Reeve
 K. Ann Renninger
 David Rettinger
 Christopher Rhoads
 Tobias Richter
 Christina Rinaldi
 Cynthia Ann Rohrbeck
 Ido Roll
 Rod D. Roscoe
 Scott Ross
 Christine M. Rubie-Davies
 Robert Rydell

 Walter C. de O. Sa
 John Sabatini
 Paul R. Sackett
 Christine Salzer
 Carol Sansone
 Varma Sashank
 Holly Schindler
 Claudia Schoene
 Johannes Schult
 Malte Schwinger
 Tina Seidel
 Corwin Senko
 Michael J. Serra
 Priti Shah
 Cynthia R. Shanahan
 David L. Share
 Amy Shelton
 Jill Talley Shelton
 Kelly Sheperd
 Georgios D. Sideridis
 Lori Skibbe
 Sheri-Lynn Skwarchuk
 Emily Slusser
 Jonathan Smallwood
 Bart Soenens
 Emily J. Solari
 Tom Southern
 Jorn R. Sparfeldt
 Ian Spence
 Jessaca Spybrook
 Jon R. Star
 Dorothy Steffler
 Hillary H. Steiner
 Ricarda Steinmayr
 Donald M. Stenhoff
 Joachim Stiensmeier-Pelster
 Gijsbert Stoet
 Martin Storksdieck
 Mari Strand Cary

 Karla K. Stuebing
 Robert H. Stupnisky
 Anna Sudkamp
 Shuyan Sun
 Rosemary Sutton

 S. K. Uma Tauber
 Florentina Taylor
 Melissa S. Terlecki
 Dave J. Theriault
 Sigmund Tobias
 Tammy Tolar
 Martin Tomaski
 Stephen Tonks
 Ulrich Trautwein
 Wendy Troop-Gordon
 Jeannine E. Turner

 Paul van den Broek
 Jacques van der Meer
 Sanne Van der Ven
 Tamara van Gog
 Maarten Vansteenkiste
 Michele V. Vecchione
 Lieven Verschaffel
 Eduardo Vidal-Abarca
 Courtney von Hippel
 Elizabeth Votruba-Drzal
 Rose K. Vukovic
 Michael Vuolo

 Jonathan Wai
 Joan M. T. Walker
 Sharon Walpole
 Lijuan Wang
 Qian Wang
 Christopher A. Was
 Hersh Waxman
 Denny Way
 Rose Mary Webb
 Mi-Young Webb
 Kristin Weingartner
 Jade Wexler
 Michelle Wilkerson-Jerde
 Gary L. Williamson
 Linda Wirthwein
 Michael B. Wolfe
 Christopher R. Wolfe
 Meng-Jia Wu
 Wei Wu
 Sascha Wuestenberg

 David Yeager
 Steven R. Yussen

 Matthias Ziegler
 Friederike Zimmermann
 Rebecca Zwick



Charles C Thomas

PUBLISHER • LTD.

P.O. Box 19265
Springfield, IL 62794-9265

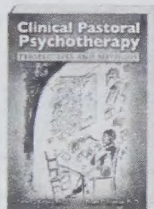
COMING SOON!

- Smith, Cary Stacy & Li-Ching Hung—**SUBCLINICAL PSYCHOPATHS: How They Adapt, Their Interpersonal Interactions with and Effect on Others, and How to Detect Them.** '13, 246 pp. (7 x 10), about \$55.95, (hard), about \$35.95, (paper), about \$35.95, (ebook).

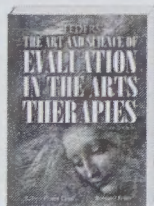
NOW AVAILABLE!



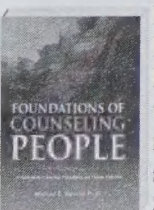
Carroll, Debbie & Claire Lefebvre—**CLINICAL IMPROVISATION TECHNIQUES IN MUSIC THERAPY—A GUIDE FOR STUDENTS, CLINICIANS AND EDUCATORS.** '13, 118 pp. (8 1/2 x 11), 11 il., \$27.95, (spiral) paper, \$27.95, (ebook).



Kaplan, Steven J. & Bruce P. Forman—**CLINICAL PASTORAL PSYCHOTHERAPY: Perspectives and Methods.** '13, 192 pp. (7 x 10), \$29.95, (paper), \$29.95, (ebook).



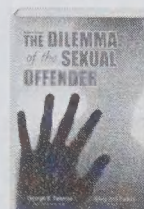
Cruz, Robyn Flaum & Bernard Feder—**FEDERS' THE ART AND SCIENCE OF EVALUATION IN THE ARTS THERAPIES: How Do You Know What's Working?** (2nd Ed.) '13, 420 pp. (7 x 10), 11 il., 5 tables, \$73.95, (hard), \$53.95, (paper), \$53.95, (ebook).



Illovsky, E. Michael—**FOUNDATIONS OF COUNSELING PEOPLE: A Guide for the Counseling, Psychological, and Helping Professions.** '13, 286 pp. (7 x 10), \$59.95, (hard), \$39.95, (paper), \$39.95, (ebook).



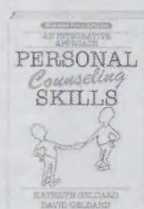
Lester, David & John F. Gunn III—**SUICIDE IN PROFESSIONAL AND AMATEUR ATHLETES: Incidence, Risk Factors, and Prevention.** '13, 262 pp. (7 x 10), 13 tables, \$58.95, (hard), \$38.95, (paper), \$38.95, (ebook).



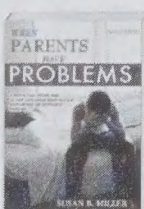
Palermo, George B. & Mary Ann Farkas—**THE DILEMMA OF THE SEXUAL OFFENDER.** (2nd Ed.) '13, 356 pp. (7 x 10), \$69.95, hard, \$49.95, paper, \$49.95, (ebook).



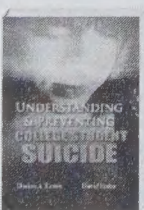
Gottlieb, Linda J.—**THE PARENTAL ALIENATION SYNDROME: A Family Therapy and Collaborative Systems Approach to Amelioration.** '12, 302 pp. (7 x 10), \$64.95, hard, \$44.95, paper.



Geldard, Kathryn & David Geldard—**PERSONAL COUNSELING SKILLS: An Integrative Approach.** (Rev. 1st Ed.) '12, 340 pp. (7 x 10), 20 il., 3 tables, \$45.95, paper.



Miller, Susan B.—**WHEN PARENTS HAVE PROBLEMS: A Book for Teens and Older Children Who Have a Disturbed or Difficult Parent.** (2nd Ed.) '12, 120 pp. (7 x 10), \$19.95, paper.



Lamis, Dorian A. & David Lester—**UNDERSTANDING AND PREVENTING COLLEGE STUDENT SUICIDE.** '11, 360 pp. (7 x 10), 21 il., 11 tables, \$69.95, hard, \$49.95, paper.

BOOK SAVINGS!

(on separate titles only)

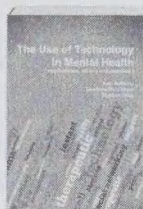
Save 10% on 1 Book !

Save 15% on 2 Books !

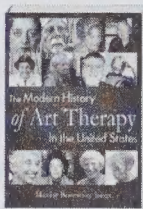
Save 20% on 3 Books !



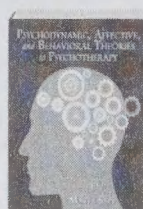
Goodman, Karen D.—**MUSIC THERAPY EDUCATION AND TRAINING: From Theory to Practice.** '11, 342 pp. (7 x 10), 3 tables, \$74.95, hard, \$54.95, paper.



Anthony, Kate, DeeAnna Merz Nagel & Stephen Goss—**THE USE OF TECHNOLOGY IN MENTAL HEALTH: Applications, Ethics and Practice.** '10, 354 pp. (7 x 10), 6 il., 5 tables, \$79.95, hard, \$54.95, paper.



Junge, Maxine Borowsky—**THE MODERN HISTORY OF ART THERAPY IN THE UNITED STATES.** '10, 370 pp. (7 x 10), 19 il., 1 table, \$77.95, hard, \$57.95, paper.



Sapp, Marty—**PSYCHODYNAMIC, AFFECTIVE, AND BEHAVIORAL THEORIES TO PSYCHOTHERAPY.** '10, 242 pp. (7 x 10), 8 tables, \$59.95, hard, \$39.95, paper.

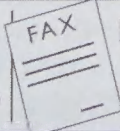


Sapp, Marty—**PSYCHOLOGICAL AND EDUCATIONAL TEST SCORES: What Are They?** '02, 204 pp. (7 x 10), 3 il., 21 tables, \$33.95, paper.

5 easy ways to order!



PHONE:
1-800-258-8980
or (217) 789-8980



FAX:
(217) 789-9130



EMAIL:
books@ccthomas.com

Web: www.ccthomas.com



MAIL:
Charles C Thomas •
Publisher, Ltd.
P.O. Box 19265
Springfield, IL 62794-9265

Complete catalog available at www.ccthomas.com or email books@ccthomas.com

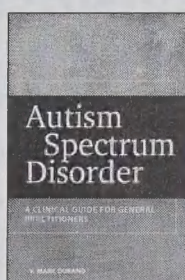
Books sent on approval • Shipping charges: \$9.75 min. U.S. / Outside U.S., actual shipping fees will be charged • Prices subject to change without notice

*Savings include all titles shown here and on our web site. For a limited time only.

When ordering, please refer to promotional code JEDP1013 to receive your discount.

NEW RELEASES

from the American Psychological Association

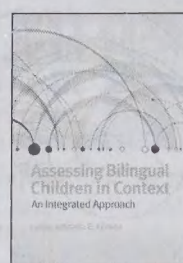


Autism Spectrum Disorder A Clinical Guide for General Practitioners

V. Mark Durand

2014. 216 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1569-0 | Item # 4317325



Assessing Bilingual Children in Context

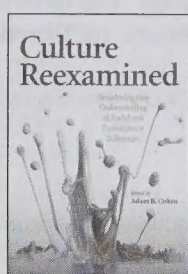
An Integrated Approach

Edited by Amanda B. Clinton

2014. 281 pages. Hardcover.

• Series: Division 16: School Psychology

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1565-2 | Item # 4317323

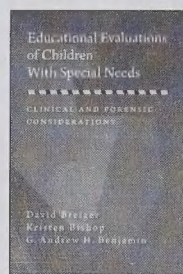


Culture Reexamined Broadening Our Understanding of Social and Evolutionary Influences

Edited by Adam B. Cohen

2014. 256 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1587-4 | Item # 4316159



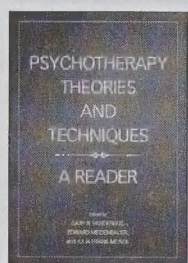
Educational Evaluations of Children With Special Needs Clinical and Forensic Considerations

David Breiger, Kristen Bishop,
and G. Andrew H. Benjamin

2014. 152 pages. Hardcover.

• Series: Forensic Practice in Psychology

List: \$59.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1575-1 | Item # 4317326



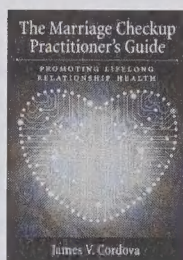
Psychotherapy Theories and Techniques

A Reader

Edited by Gary R. VandenBos, Edward
Meidenbauer, and Julia Frank-McNeil

2014. 368 pages. Paperback.

List: \$34.95 | APA Member/Affiliate: \$29.95
ISBN 978-1-4338-1619-2 | Item # 4317329



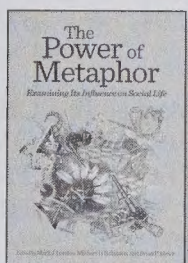
The Marriage Checkup Practitioner's Guide

Promoting Lifelong Relationship Health

James V. Cordova

2014. 264 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1552-2 | Item # 4317319

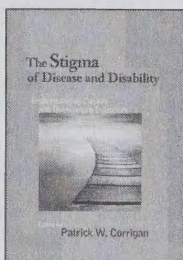


The Power of Metaphor Examining Its Influence on Social Life

Edited by Mark J. Landau, Michael D.
Robinson, and Brian P. Meier

2014. 304 pages. Hardcover.

List: \$79.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1579-9 | Item # 4318123



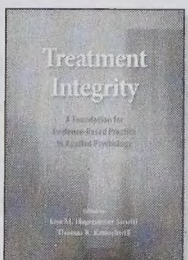
The Stigma of Disease and Disability

Understanding Causes
and Overcoming Injustices

Edited by Patrick W. Corrigan

2014. 312 pages. Hardcover.

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1583-6 | Item # 4318124



Treatment Integrity

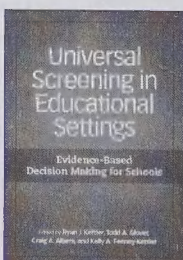
A Foundation for Evidence-Based
Practice in Applied Psychology

Edited by Lisa M. Hagermoser Sanetti
and Thomas R. Kratochwill

2014. 320 pages. Hardcover.

• Series: Division 16: School Psychology

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1581-2 | Item # 4317327



Universal Screening in Educational Settings Evidence-Based Decision Making for Schools

Edited by Ryan J. Kettler, Todd A. Glover,
Craig A. Albers, and Kelly A. Feeney-Kettler

2014. 328 pages. Hardcover.

• Series: Division 16: School Psychology

List: \$69.95 | APA Member/Affiliate: \$49.95
ISBN 978-1-4338-1550-8 | Item # 4317318



AMERICAN PSYCHOLOGICAL ASSOCIATION

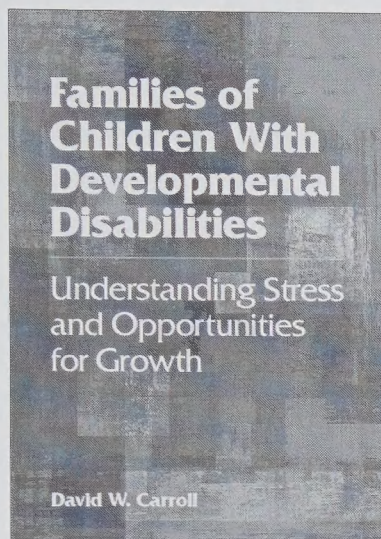
TO ORDER: 800-374-2721 • www.apa.org/pubs/books

AD2297

FAMILIES OF CHILDREN WITH DEVELOPMENTAL DISABILITIES

Understanding Stress and Opportunities for Growth

David W. Carroll



Parents of children with disabilities confront a number of challenges and may be at risk for depressive or trauma-related symptoms. Changes in family roles and routines can cause stress for parents, siblings, and extended family alike as they confront multiple issues, including behavioral problems and frequent healthcare needs. Despite such challenges, many families derive a sense of meaning from facing their difficulties in a positive way. This book surveys the most recent empirical research on families of children with disabilities and provides guidelines and strategies for the developmental and family psychologists who support these clients.

The book follows a developmental progression, first examining the immediate effects that a child's disability can have on his or her family and looking at the changes that occur as the child grows and faces new challenges. In doing so, the author examines studies employing a variety of methodologies, including quantitative research, meta-analyses, and qualitative methods such as narrative analysis. The book also describes cognitive behavioral interventions and

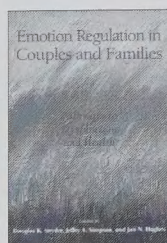
programs that train parents to more effectively manage child behavioral problems and thereby improve family well-being. 2013. 240 pages. Hardcover.

List: \$59.95 | APA Member/Affiliate: \$49.95 | ISBN 978-1-4338-1329-0 | Item # 4316155

CONTENTS:

Chapter 1: Introduction | Chapter 2: Initial Experience and Reactions | Chapter 3: Stress, Coping, and Growth | Chapter 4: Family Change and Reorganization | Chapter 5: Medical Issues and Medical Professionals | Chapter 6: Special Education, Inclusion, and Advocacy | Chapter 7: Social Exclusion and Social Support | Chapter 8: Developmental Disabilities Through the Lifespan | Chapter 9: Life Challenges and Life Stories | Chapter 10: Death and Bereavement | Chapter 11: Clinical Implications | Chapter 12: Conclusions and Future Directions

ALSO OF INTEREST



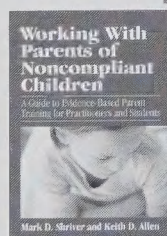
Emotion Regulation in Couples and Families *Pathways to Dysfunction and Health*

Edited by Douglas K. Snyder,
Jeffry A. Simpson,
and Jan N. Hughes

2006. 332 pages. Hardcover.

List: \$29.95 | APA Member/Affiliate: \$24.95
ISBN 978-1-59147-394-7 | Item # 4318032

Working With Parents of Noncompliant Children



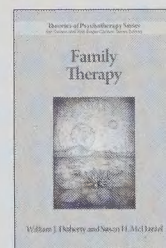
A Guide to Evidence-Based Parent Training for Practitioners and Students

Mark D. Shriver
and Keith D. Allen

2008. 304 pages.
Hardcover.

• Series: Division 16: School Psychology

List: \$39.95 | APA Member/Affiliate: \$34.95
ISBN 978-1-4338-0344-4 | Item # 4317155



AVAILABLE ON AMAZON KINDLE® **Family Therapy**

William J. Doherty
and Susan H.
McDaniel

2010. 125 pages. Paperback.

• Series: Theories of
Psychotherapy Series®

List: \$24.95 | APA Member/Affiliate: \$24.95
ISBN 978-1-4338-0549-3 | Item # 4317202



AMERICAN PSYCHOLOGICAL ASSOCIATION

APA BOOKS ORDERING INFORMATION: 800-374-2721 • www.apa.org/pubs/books

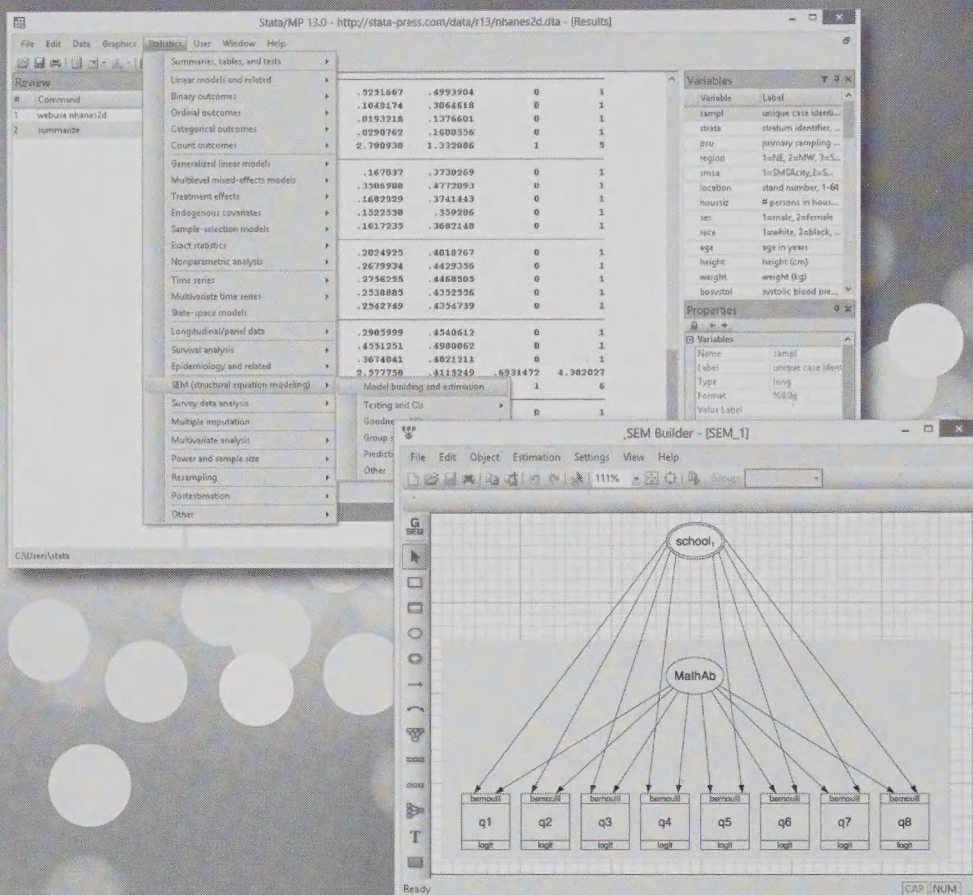
In Washington, DC, call: 202-336-5510 • TDD/TTY: 202-336-6123 • Fax: 202-336-5502

In Europe, Africa, or the Middle East, call: +44 (0) 1767 604972

AD2282

STATA[®] 13

Statistics • Graphics • Data management



Stata 13 provides everything you need for statistical analysis, data management, and graphics in one integrated package.

13 is now your lucky number.

stata.com/edu13

New in Stata 13

Effect sizes

Comparisons	ANOVA
<i>t</i> -tests	Cohen's <i>D</i>
Hedges's <i>G</i>	Glass's Δ
η^2	ω^2

Generalized SEM

Multilevel and hierarchical models

More outcomes:

- » Binary (logit and probit)
- » Counts (Poisson and negative binomial)
- » Categorical (multinomial logit)
- » Ordered (ordered logit and probit)

More models:

- » CFA with binary, count, and ordinal measurements
- » Multilevel CFA
- » Multilevel mediation
- » Item response theory (IRT)
- » Latent growth curves with repeated measurements of binary, count, and/or ordinal responses

Treatment effects

- » Inverse probability weights (IPW)
- » Regression adjustment
- » Propensity-score matching
- » Covariate matching
- » Doubly robust methods

Multilevel hierarchical models

Linear	Logistic
Probit	Ordered logistic & probit
Poisson	Negative binomial

